HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Information and Natural Sciences

Lauri Kovanen

Structure and dynamics of a large-scale complex social network

Master's thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology in the Degree Programme for Engineering Physics.

Espoo, 10.03.2009

Supervisor : Professor Kimmo Kaski
Instructor : Dr.Tech. Jari Saramäki

Informaatio- ja luonnontieteiden tiedekunta

Tekijä:	*Lauri Kovanen*
Koulutusohjelma:	*Teknillisen fysiikan koulutusohjelma*
Pääaine:	*Informaatiotekniikka*
Sivuaineet:	*Laskennallinen tekniikka, Kansantaloustiede*

Työn nimi:	*Suurikokoisen kompleksisen sosiaalisen verkoston rakenne ja dynamiikka*
Title in English:	*Structure and dynamics of a large-scale complex social network*
Professuurin koodi ja nimi:	*S-114 Laskennallinen tekniikka*
Työn valvoja:	*Professori Kimmo Kaski*
Työn ohjaaja:	*TkT Jari Saramäki*

Tiivistelmä:

Suurten verkostomuotoisten tietoaineistojen tutkimus on lisääntynyt valtavasti viimeisen vuosikymmenen aikana. Tämä johtuu sekä tietotekniikan kehittymisestä, joka mahdollistaa suurten aineistojen käsittelyn, myös sähköisen viestinnän lisääntymisestä, josta on huomattavan helppoa kerätä tarkkaa tietoa.

Tässä työssä tutkitaan yhden eurooppalaisen matkapuhelinoperaattorin laskutukseen perustuvaa aineistoa. Alustavan analyysin perusteella voidaan sanoa, että aineisto on hyvin heterogeenistä: esimerkiksi yksilöiden käyttäytyminen vaihtelee paljon, samoin kuin kommunikointi eri vuorokaudenaikoina ja viikonpäivinä.

Tarkemmin paneudutaan kahteen uuteen aiheeseen. Ensimmäisenä tutkitaan matkapuhelinviestinnän vastavuoroisuutta. Tasapuolisuus osoittautuu suhteellisen harvinaiseksi; toinen osapuoli soittaa usein huomattavasti enemmän kuin toinen. Tämä ei kuitenkaan johdu vain siitä, että toiset ihmiset ovat ylipäätään aktiivisempia, vaan tasapuolisuuden puute vaikuttaa olevan suhteen ominaisuus. Lisäksi voimme todeta, että aktiivisempi osapuoli on usein se, jolla on enemmän tuttavia.

Toisena aiheena tutkitaan kausaalisuutta. Onko mahdollista erottaa, kuinka suuri osa soitetuista puheluista johtuu aiemmista vastatuista puheluista? Tähän kysymykseen on mahdotonta vastata minkään yksittäisen puhelun osalta, koska emme tunne puheluiden sisältöä, mutta voimme silti sanoa jotain keskiarvoista. Osoittautuu esimerkiksi, että jos vastattu puhelu saa aikaan uuden puhelun, se soitetaan keskimäärin 25 sekunnin kuluttua edellisen puhelun loppumisesta. Ongelmaksi muodostuu kuitenkin se, kuinka erottaa kausaalisuus korrelaatiosta.

Sivumäärä: *72* Avainsanat: *kompleksiset systeemit, kompleksiset verkot, sosiaaliset verkot, kausaalisuus, vastavuoroisuus, matkaviestintä*

Täytetään tiedekunnassa

Hyväksytty: Kirjasto:

Faculty of Information and Natural Sciences

Author:	*Lauri Kovanen*
Degree programme:	*Degree Programme for Engineering Physics*
Major subject:	*Computer and Information Science*
Minor subjects:	*Computational Engineering, Economics*
Title:	*Structure and dynamics of a large-scale complex social network*
Title in Finnish:	*Suurikokoisen kompleksisen sosiaalisen verkoston rakenne ja dynamiikka*
Chair:	*S-114 Computational engineering*
Supervisor:	*Professor Kimmo Kaski*
Instructor:	*Dr.Tech. Jari Saramäki*

Abstract:

Research on large-scale complex social networks has increased considerably during the last decade. This has two basic reasons: the growth of computational power allows the study of ever larger data, and secondly the prevalence of electric communication makes it easier to collect data on a large number of people.

This thesis studies a data set based on the billing information of one European mobile phone operator. Preliminary analysis reveals that the data is very heterogeneous: there are large variations between people, as well between different times of day and different weekdays.

The thesis concentrates on two novel features. First we'll study the reciprocity of mobile phone communication. It turns out that evenness is quite rare; it's relatively common that one party is responsible for a large fraction of the calls. This is not only a result of the heterogeneity in human activity, but the unevenness seems to be a property of the relationship itself. Furthermore, the more active party appears to be the one with more acquaintances.

The second feature concerns causality. Is it possible to tell how large a fraction of calls are caused by earlier calls? It is impossible to give an answer in respect to any single call, since we do not know the contents of the calls, but we can still say something about averages. It turns out that if for example an incoming call induces a new call, that call takes place on average 25 seconds after the previous call ends. A large problem in this study is distinguishing causality from correlation.

Number of pages: *72*

Keywords: *complex systems, complex networks, social networks, causality, reciprocity, mobile communication*

Faculty fills

Approved: Library code:

Preface

This work was carried out in the Department of Biomedical Engineering and Computational Science at Helsinki University of Technology.

I wish the express my gratitude to the whole research group: my instructor Dr.Tech. Jari Saramäki for his insights and comments during the work, professor Kimmo Kaski for creating such a great place to work at, Mikko Kivelä for a multitude of useful and practical comments, and all other people at the lab who have made this work much more enjoyable than one could hope for.

I also wish to thank Albert-László Barabási and Julian Candia at Notre Dame University for providing the mobile phone data set, without which this Thesis would not have been possible.

Finally, I wish to thank my family for never telling me what I should do, and my lovely girlfriend Johanna for her most delightful support.

Otaniemi, March 10$^{\text{th}}$, 2009

Lauri Kovanen

Contents

Chapter 1

Introduction

The recent decade has seen a vast increase in research of complex social networks. The increase of computational power has allowed the analysis of ever larger networks, while the constant evolution of digital telecommunication systems has made it feasible to collect gigantic yet precise data sets about human relationships. The largest network studied so far can easily be called "planetary"[1], as it describes the communication patterns between 180 million people all around the world [1]. This is a massive change compared to for instance the 34 members of Zachary's karate club [2], a network data extensively studied by sociologists ever since its publication in 1977.

Quite naturally, networks with millions of nodes can not be treated with the same accuracy as networks with only a dozen or so people. It becomes unfeasible to concentrate on the characteristics of single individuals, firstly because we do not have accurate information on their motives, skills, habits, dreams, and personal histories, and secondly because analysing millions of individuals with such accuracy is too large of a task. Instead, our analysis will be based on distributions and statistical properties. Besides, many of the phenomena studied here are only visible when we take a step back and look

[1] "Global", however, would not be a correct term to use. Because the aforementioned data consists of instant messaging communications, it quietly excludes the most populous areas on our planet, like many African countries and rural China.

at the bigger picture.

If the reader is worried that aggregating people into distributions and placing numbers on their properties somehow underestimates the inherent uniqueness of individuals, I can assure you there is no reason for such concern. Quite the contrary; again and again we'll run into distributions with a variance so large that there can not be said to be a typical individual. This is a strong demonstration of the existence of uniqueness, no matter where we look.

1.1 The aim of this thesis

This thesis will concentrate more closely on two little studied properties of complex social networks. The first property (Chapter 4) concerns the static structure of the network, and discusses the reciprocity of the edges. The purpose is to study how uneven or equal are the relations between any two individuals. Note that here the word 'static' means that the analysis doesn't pay attention to the time-varying nature of the network. The opposite is 'dynamic', implying that temporal changes in the network are taken into account.

The second property of interest is the causality of calling behaviour (Chapter 5). We'll apply a method originally introduced to the study of neural cells in an attempt to find out whether arriving calls can be said to cause new calls. The study of causality quite naturally falls into the 'dynamic' category.

The selection of these two properties as the focus of this thesis might seem fairly haphazard. To some extent, this is true. The field of social network research has grown to encompass such a large range of different subjects and methods that it's impossible to cover them all in only one thesis. On the other hand, the chosen properties have both been covered very scarcely in the current social network literature, and therefore this choice takes this thesis to the front line of research. Secondly, the choice can be justified by purely practical reasons — the data set utilized has plenty of relevant and

accurate information for the study of these two properties.

Before we begin the actual analysis, Chapter 2 will go through the necessary concepts and definitions to introduce the reader to the world of networks. These concepts are then put to use in Chapter 3 where we introduce the data that is used throughout the analysis.

Finally Chapter 6 gleans together all the bits of information found on the way to create a mental picture of the gigantic network. After all, the simple question we are after is "What does a large social network look like?"

Chapter 2

About networks and complexity

This chapter will define the necessary concepts for understanding graphs and in particular their application to the study of social networks. If you're already familiar with the study of social networks you may skip straight to Chapter 3.

2.1 Networks

A network (or *a graph*) is mathematically defined as a pair $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of *nodes* (or *vertices*), connected by *edges* $\mathbf{E}$. Each edge connects two nodes,[1] thus an edge $e \in \mathbf{E}$ that connects the node $i \in \mathbf{V}$ to node $j \in \mathbf{V}$ may be written as $e_k = (i, j)$.

In a *directed network* the two edges e_{ij} and e_{ji} are different objects, whereas in an *undirected network* the ordering of the end points of the edge has no relevance: (i, j) and (j, i) refer to the same edge. A food web is an example of a naturally directed network: the nodes represent different species and the existence of edge (i, j) means that species i eats species j. Collaboration

[1] In fact in *hypergraphs* a single edge may connect more than two nodes. However, hypergraphs are not relevant in this thesis, and will not be discussed further.

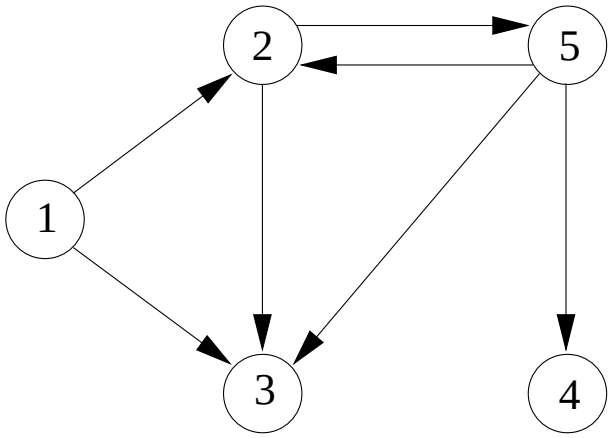


Figure 2.1: A directed network with $\mathbf{V} = \{1, 2, 3, 4, 5\}$ and $\mathbf{E} = \{(1, 2), (1, 3), (2, 3), (2, 5), (5, 2), (5, 3), (5, 4)\}$.

networks are undirected: each node represents one person and the existence of edge $(i, j) = (j, i)$ means that i and j have collaborated.

We use $N = |\mathbf{V}|$ for the number of nodes and $L = |\mathbf{E}|$ for the number of edges. Assuming the network has no self-loops (edges of the form (i, i)), the maximum number of edges in a directed network is $N(N-1)$ and in an undirected network $N(N-1)/2$. The *density* of a network if defined as the proportion of all possible edges that exist. For a directed network this equals $\frac{L}{N(N-1)}$.

A synonym for a network is *graph*. The word 'network' has however become the standard term in physics, while 'graph' is more common in purely mathematical contexts. 'Network' will also be the preferred term in this thesis.

2.1.1 Unweighted networks

An unweighted network is determined by the node set and the existence or non-existence of edges between the nodes. One way to write down the full description of any arbitrary network is by using an *adjacency list*. The adjacency list has one entry (one row in written text) for each node in the

network, consisting of the identifier of the node followed by a list of nodes connected to it. For example, the adjacency list of the network in Figure 2.1 is

1: 2, 3
2: 3, 5
3:
4:
5: 2, 3, 4

In this example the nodes are identified with positive integers running from 1 to $N = 5$. The labels of the nodes need not be integers, or even numbers, and in some examples in this thesis letters will be used instead. Integers are however the most natural choice when the nodes are initially anonymous, and by far the easiest choice when the network is handled with a computer.

Alternatively the network may be represented with an *adjacency matrix* A, where $A_{ij} = 1$ if the edge (i, j) is present in the network, and otherwise $A_{ij} = 0$. For the example network in Fig 2.1 we have

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} .$$

For undirected networks the adjacency matrix is always symmetric. Note that the diagonal contains only zeros since our model network does not contain any self-loops.

2.1.2 Weighted networks

In unweighted networks the edge are binary: they either exist or don't. In many applications this binary representation is not realistic. The above in-

troduction to unweighted networks may be extended by assigning a weight w_{ij} to each edge. Much like the identifiers of nodes, the weight could be any object, a piece of text or a vector, but in most networks the weights are either integers or real numbers.

The elements of the adjacency matrix are now the weights: $A_{ij} = w_{ij}$. If the weights are strictly positive numbers, we can use zero as the weight of non-existent edges. If this is not the case, we'll need one matrix for the weights and another for the existence of edges. In the adjacency list representation the list of neighbouring nodes must be accompanied with a list of edge weights.

While social networks may be represented by a unweighted network (an edge exists if two people know each other) they benefit greatly from the weighted representation — people may have hundreds of acquaintances, but only a handful of them are significant in everyday life. Depicting human relationships simply on the scale "existing–non-existing" is not very useful.

2.1.3 Additional concepts and definitions

Here we introduce some useful concepts and definitions for the analysis of networks.

Degree and strength

Degree of a node in an undirected network is the number of incident edges, commonly denoted by k. Because each edge has two ends and therefore contributes to the degree of two nodes, the average degree of a network is

$$\overline{k} = \frac{2|E|}{|V|} = \frac{2L}{N} \quad . \tag{2.1}$$

An analogy to the degree in undirected weighted networks is the *strength* of a node, defined as the sum of the weights of incident edges:

$$s_i = \sum_j w_{ij} \quad . \tag{2.2}$$

In directed networks the in-degree and out-degree (or in-strength and out-strength in weighted networks) are in not equal in general. However, since each directed edge contributes equally to the total in-degree and the total out-degree, the average in- and out-degrees are equal. The same is true for average in- and out-strengths.

Path and the geodesic distance

A path from node i to node j is an alternating succession of nodes and edges,

$$v_0 e_0 v_1 e_1 v_2 e_2 \dots e_{l-1} v_l, \ v_k \in \mathbf{V}, \ e_k \in \mathbf{E} \quad ,$$

such that $v_0 = i$ and $v_l = j$. Since any edge or node can appear an unlimited number of times in the path, the number of different paths between any two nodes is always either zero or infinite. *A loop* is a path that starts from and ends to the same node, that is, $v_0 = v_l = i$.

The geodesic distance d_{ij} is the length of the shortest path between nodes i and j, where the length is defined as the number of traversed edges; in weighted networks the path length may also be measured by the sum of the edge weights. When the edge weights are non-negative the shortest path can not have loops; if it did, we could construct a shorter path by removing the loop. However, the shortest path may not be unambiguous, since there can be several paths with the same total length d_{ij}.

Assortativity

A network is said to be *assortative* if the average degree of the neighbourhood grows with the node degree. This can be measured with the assortativity coefficient, defined in [3] as the Pearson correlation coefficient of the degrees of adjacent nodes. A more thorough picture of assortativity can be gained by plotting the average degree of neighbours as a function of node degree. Assortative networks are characterised by a rising curve.

Clustering coefficient

The clustering coefficient of node i is defined as

$$c(i) = \frac{t_i}{k_i(k_i - 1)/2} \quad , \tag{2.3}$$

where k_i is the degree of node i and t_i is the number of edges among the neighbours of node i. The clustering coefficient $C(i)$ can be interpreted as the probability that two randomly chosen neighbours of node i are connected. The clustering coefficient is often averaged over all nodes of the same degree, defining

$$c_k = \frac{\sum_{\deg(i)=k} c(i)}{|\{i \in \mathbf{V} | \deg(i) = k\}|} \quad . \tag{2.4}$$

c_k can be thought as the estimate for the probability that two randomly chosen neighbours of a node with degree k are connected.

The clustering coefficient can not be unambiguously extended to weighted or directed networks. Ref. [4] compares 4 definitions of weighted clustering coefficients and concludes that "It is clear ... that there is no ultimate formulation for a weighted clustering coefficient." On the other hand, [5] uses four different directed clustering coefficients to define a clustering signature of a directed network. As clustering behaviour is not a central concept in this thesis, we'll restrain from using the weighted and directed clustering coefficient and just stick to the classical definition of Eq. (2.3).

Another unconnected problem with the clustering coefficient is that it correlates with assortativity. This is especially pronounced in disassortative networks: when all neighbours of a large-degree node have a very small degree, it naturally follows that the clustering coefficient of the large-degree node must be small, as there can not be many edges between the neighbours. A similar reasoning is valid in most assortative networks also because the neighbours of the largest-degree nodes have much smaller degrees in comparison.

To remove the correlation between the assortativity and clustering coefficient, [6] defines *the uncorrelated clustering coefficient* as

$$\tilde{c}(i) = \frac{t_i}{\omega_i} \quad , \tag{2.5}$$

where ω_i is the maximum number of edges possible between the neighbours given their degrees. Note that because $\omega_i \leq k_i(k_i - 1)/2$, $\tilde{c}(i) \geq c(i)$. The limit $\tilde{c}(i) = c(i)$ is reached when all neighbours are fully connected to each other. A reasonably fast algorithm for calculating ω_i is given in [6].

2.2 Social networks

Social networks are graphs where the nodes correspond to people (each node represents one individual) and the edges correspond to some relation between the two people. Social networks are subjective in the sense that the existence of edges depends on our definition of the relation. A good breakdown of possible relations is given in [7], where the relations are divided into four groups:

Similarities of location, membership or other attribute (gender, opinion, etc.)

Social relations such as kinship or other role (mother of, son of, boss of, etc.), affective (likes/hates) or cognitive (knows, knows about, etc.)

Interactions: Helped, collaborated, had sex with, talked to, etc.

Flows of information, beliefs, resources, etc.

This classification gives some perspective about the subject of this thesis, where the data consists of interactions taking place over mobile phones. There are most certainly many aspects of social life that are missing from the data, even if we were to make the (audacious) assumption that mobile phone communication correlates with social proximity.

Following the discussion about different relations it is now obvious that it not possible to construct the 'true' social network. Many different kinds of social networks have already been studied, and the choice of relation very often depends on both the purpose of the study and the data currently available. Examples of some of the most widely studied social networks in networks literature include

Scientific collaboration networks where the nodes are researchers, often representing one field of study. There is an edge between two scientists if they have collaborated to write an article. The edge weight can be used to denote the number of common articles. [8, 9]

Sexual contact networks where two people (two nodes) are connected if they have had a sexual relationship. [10]

Instant messaging networks where each node represents one user ID of the instant messaging system (most often each user ID is used by only one person), and two user IDs are linked if they've had a conversation. The edge weight could denote either the number of separate conversations, the number of individual messages or the total length of the conversation. [1]

Mobile phone networks where a (directed) edge (i, j) exists if person i has called to person j. The weight can denote the total number of calls or the total duration of calls [11, 12, 13]. Mobile phone networks are also the subject of this thesis.

Note that some networks are inherently undirected, as is the case with the collaboration and sexual contact networks, while the instant messaging and mobile phone networks are directed.

Social network were extensively studied by sociologists for over hundred years before physicists caught on and started applying methods earlier used only in statistical physics. This history is still eminently visible even today in the differing research frameworks applied implicitly by sociologists and physicists. Physicists, for instance, have not been too worried about the differences between the types of edges, applying the same methods to sexual contact networks, collaboration networks and even to protein interaction networks. Sociologists are also more keen on stressing the importance of external variables, such as age and gender, and their effect on the network structure. While these different frameworks hinder the exchange of knowledge between the two factions, it may also produce fresh and surprising discoveries. It is hard to argue with the conclusion of [7] that there are undoubtedly many lessons to be learned from the other.

While it is obvious that the networks of scientific collaborators and sexual contacts do not in general overlap[2], it is so much more surprising to notice that social networks with varying definitions for edges share many common characteristics.

Dynamic Networks

Obviously social networks are not static objects. If we look at a social network at any time instant, we can expect nearly all nodes to be replaced roughly once every 100 years simply because of the limited life-span of our species. In addition, the social contacts we interact with change from year to year, month to month and even from hour to hour — for instance, we talk with different people during working hours than during weekends. Since some social network data also include a temporal component, such as the sending times of e-mail

[2]Of course, workplace romances do happen even among researcher, probably the most famous example being that of Marie and Pierre Curie.

messages, we are able to study also the temporal evolution and processes taking place in networks. Let's take a look at some recent advancements in this area.

The article by Palla, Barabási and Vicsek [14] in 2007 studies the temporal evolution of communities[3] in two social networks, a collaboration network with 30000 authors and a mobile phone network of over 4 million users. The study finds a similar behaviour in both networks: small communities have a longer life-span if the members stay the same, while large communities last longer if the members are changed continuously. These correspond roughly to two different types of durable communities: fixed friendship networks and large institutions.

One central idea in the study of network dynamics is the concept of *time-respecting paths*, that is, directed paths where each edge has to be newer than the previous one. For example, the shortest time-respecting path in an e-mail network is the fastest way information (or a furious computer virus) could spread. Using this idea, Ref. [15] studies the temporal reachability of nodes in e-mail networks and finds that they have a dense core surrounded by a sparser periphery. Because e-mails are often sent in bursts, and because there are strong daily and weekly patterns, the aggregated static network would not give the same insight about the information flow.

Ref. [16] uses a similar concept of *vector clocks* to study the spread of information. It first defines *information latency* $t - \phi_{j,t}(i)$ as the amount of time the node j is out-of-date about i at time t. Here $\phi_{j,t}(i)$ is the largest time $t' < t$ so that a message sent by i at time t' reaches j by time t. Vector clock ϕ_j is the collection of information latencies with respect to each other node, calculated at each time instant. The study finds, among other interesting results, that the observed e-mail communication pattern is a nearly optimal compromise (with respect to information flow) between concentrating all traffic on most important edges and levelling the traffic evenly on all edges.

[3]Dense groups of nodes; see 2.2.4.

Vassilis Kostakos [17] uses very similar notation to define *temporal graphs* where a path between two nodes is automatically the time respecting path. The article examines two different data sets, an email corpus and a data on people's face to face encounters, and concludes that even though the two data sets are structurally similar in terms of static distributions, the temporal behaviour varies greatly. Also, the people who received information quickly were not necessarily good at spreading information.

While there isn't yet a large number of articles on network dynamics, the studies carried out thus far do agree on one fundamental issue: social networks are definitely not static structures, and treating them as such might take us down the wrong road. This ubiquitous observation and the small number of published articles hint that the study of temporal evolution and dynamic processes in social networks have a lot of untapped potential.

2.2.1 Clustering and assortativity

Social networks are known to have both high clustering and high assortativity. These properties have very intuitive interpretations.

High clustering means that any two friends of mine are very likely to know each other. One can think of several reasons why this is so. First of all, we do not usually meet other people entirely randomly — instead, we often get introduced to a new person by a friend, or acquaint two friends ourselves, completing a triangle of acquaintances in both cases. Furthermore, even if one does meet new people at "random", those people nearly always have something in common with us (a hobby, same workplace or university etc.), increasing the likelihood that some old friend already knows the new acquaintance.

Another universal feature of social networks is high assortativity. This means simply that popular people have, on average, more popular friends than those with only few friends. Assortativity can be thought as a consequence of a more general social phenomenon called *homophily*, which states that two people

are more likely to know each other when they are alike.

2.2.2 The small world property

Several more complex features of social networks have already been experimentally verified. Possibly the most famous one is the "six degrees of separation", initially introduced by the Hungarian novelist Frigyes Karinthy in 1929. The original version actually claims only *five* degrees of separation, with the meaning that any two people in the world are connected by a chain of no more than five acquaintances.

The first attempt to verify this claim was made by Stanley Milgram in 1960's. Jeffrey Travers and Stanley Milgram [18] describe an experiment where 296 individuals in Nebraska and Boston are asked to reach a target person in Massachusetts by sending a letter to a personally known acquaintance thought to be closer to the target. While only 64 chains (21.6 %) reached the target, the mean number of intermediaries for these chains was 5.2.

Even though the method of the study leave much to hope for, it is clear that the number of intermediaries does not grow linearly with the population size. The 'six degrees of separation' is a popular term for *the small world property*: in social networks, the average geodesic distance grows logarithmically with network size.[4] Note that this claim is not as strong as the original idea presented by Karinthy. We are not saying that *all* people are connected by at most five acquaintances, but that this is (roughly) the average value.

Although the prevalence of short paths is surprising at first, the emergence of the small world property has been explained with rather simple models. The famous article by Duncan Watts and Steven Strogatz [19] describes a simple 1-dimensional lattice that exhibits the small world phenomenon when only a small number of random edges are added. Watts and Strogatz also define 'small-world networks' as networks with both short path lengths and

[4]Infinite geodesic distances, which correspond to paths between nodes in unconnected components, are naturally excluded from the average.

high clustering. Indeed, this characterisation covers most natural networks, and such networks can be said to be somewhere between entirely random networks (with small path lengths but low clustering) and regular networks (high clustering but large path lengths). The small world property is in fact so ubiquitous that a network without it would seem very peculiar.

2.2.3 Fat-tailed distributions

Most people are familiar with the bell-shaped Gaussian curve and with the "fact" that the distributions of many human characteristic, such as height and weight, may be modelled by this particular curve.[5] What is not so well know is the universality of *fat-tailed distributions*, very broadly defined as distributions where the large variance is caused by infrequent extreme deviations, as opposed to a large number of small deviations.

This definitions encompasses a vast number of different distributions. One of the most common ones is the *power law* $p(x) \sim x^{-\alpha}$, $\alpha > 1$, the cumulative distribution of which is $\mathrm{P}(X > x) \sim x^{1-\alpha}$. The power law has two interesting special cases. If $1 < \alpha < 2$ both the mean and the variance of the distribution are infinite, and if $2 < \alpha < 3$ the mean is finite but variance is infinite.

While the idea of infinite variance might seem mindboggling at first, fat-tailed distribution are in fact quite intuitive, even for those who have never heard the term. Imagine that someone told you that at this very moment everyone in the Helsinki Central Railway Station has income of over 1000 euros. If you now had to guess the income of first person you run into, what would you say? Anything between 1500 and 5000 euros would probably be a reasonable guess. Next imagine being a guest speaker at the Finnish Millionaires' Club, where all members are required to have an income of over million euros. What would you now guess is the income of the first person you encounter? Around 2 million euros?

[5] In fact the distribution of heights fits equally well to both the normal and the log-normal distributions [20]; normal distribution might not be so normal after all.

This is the very essence of power laws. They are *scale free*: If, for example, among all those earning over 1000 euros 60 % make more than 2000 euros, then among those with income over 1 million euros the same 60 % make more than 2 million. Mathematically this may be written as

$$\frac{\mathrm{P}(X > 2000)}{\mathrm{P}(X > 1000)} = \frac{2000^{-(1+\alpha)}}{1000^{-(1+\alpha)}} = 2^{-(1+\alpha)} = \frac{2e6^{-(1+\alpha)}}{1e6^{-(1+\alpha)}} = \frac{\mathrm{P}(X > 2e6)}{\mathrm{P}(X > 1e6)}$$

with $\alpha = 1 - \log_2 0.6 \approx 1.737$. There is no fixed scale, and if you followed the above example, you probably didn't even notice changing the scale. But at the railway station you were thinking about thousands, and at the millionaires club in millions of euros.[6]

Degree and strength distributions of many complex networks are famous for fat tails, and the interpretation of these distributions is similar to the example presented above.

2.2.4 Communities

A community in a network is usually loosely defined as a group of nodes with more edges within the group than between a member of a group and a node outside the group. This definition is obviously quite vague, and there is no consensus on the exact definition of communities in networks. Most algorithms for finding the optimal set of communities implicitly define a community by the very algorithm. However, many studies have shown that no matter what the exact definition is, social networks have a rich community structure.

A little bit of thought reveals the problem with recognizing communities in social networks. People naturally belong to several communities corresponding to the groups they belong to, such as people they work with, friends

[6]Incomes are used in the example only to give an intuitive interpretation for power laws. In reality, a study of U.S. taxation data finds that incomes *do not* follow a power law, except for the richest 1-3 % [21]. Most incomes reside in a more equal exponential distribution.

from childhood, student buddies, family, hobby groups, people met while living abroad, etc. Thus the communities necessarily overlap. The sizes of the communities also span over several orders of magnitude. We may observe anything from small, tight communities consisting of only a few individuals (e.g. families) to enormous communities with millions of individual (e.g. countries or language groups). A good example of the latter one is presented in [11], where the three language groups of Belgium (French, Flemish and the bilingual community) are shown to agree with the communities of a mobile phone network.

2.3 Complexity

Now that the basics of social networks are covered, it is time to discuss that little buzzword in the title of this thesis. Large-scale networks containing anything between thousand to several million nodes are commonly called *complex networks*. What exactly is the difference between 'network' and 'complex network'? What does it mean that a network is complex?

The short and practical answer is: not much. For the most part, this terminology serves the purpose of distinguishing one field of research from many others dealing with or making use of graphs and networks.

Of course, every short answer requires a matching long answer. Wiktionary, the wiki-based open content dictionary, gives two applicable definitions for adjective *complex*:

1. Made up of multiple parts; intricate or detailed.
2. Not simple or straightforward.

Obviously networks are made up of multiple parts, nodes and edges, but it wouldn't be worthwhile to rub this in with an extra term.

The second definition is more to the point but still a bit vague. Despite the fact that we often have perfect knowledge of the parts of a particular network, its large-scale structure and dynamics are not instantly obvious. A good and simple example of this is the emergence of the small-world property from a regular graph after only few random rewirings (see 2.2.2). This also illustrates another way of understanding complexity, as a kind of middle ground between completely random and completely regular systems.

Another perspective to complexity is given by [22], which makes a distinction between *complicated systems*, such as airplanes, and a *complex systems*, such as large-scale networks. A complicated system may also consist of a huge number of parts, but its operation is always well known, computable and predictable. Moreover, the behaviour of a complicated systems is perfectly known once we understand all of its subsystems. Also, failure a single part (say, the aileron of an aircraft) can have a devastating effect on the whole system.

Complex systems, however, have no useful division into parts. For example, the natural components of a network are the nodes and edges, but such a decomposition makes the network itself disappear! In this sense complex systems can be said to have *emergent properties*: it is in necessary to adopt a different point of view for modelling the whole than when modelling the parts.

Note that this does not mean that it would somehow be impossible to derive the properties of the complex system from the properties of it's components. There is no magic in emergence. For example, it is possible, at least in theory, to calculate the phase transitions of water by starting from the properties of water molecules. However, the calculations might be too burdensome to carry out in practice.

Social networks in particular are difficult to analyse by trying to understand the components alone: we do not have sufficient knowledge of all components. If we were to model the workings of a social network bottom-up, we would need to know the thoughts and capabilities of every person involved. This is

a formidable task for many reasons: people are generally not able to predict even their own future behaviour, nor would they report it accurately if asked, even if we could ask everyone.

Complexity, understood along the treatment above, is not so much an intrinsic property of the system but a way to look at and analyse it. It is a useful approach with social phenomena because it allows us to carry out a study and make conclusions about the whole *with very little knowledge of the inner workings of single components*, instead exploiting the information available about the interactions of the parts.

Chapter 3

Mobile phone data

This chapter will introduce the mobile phone data used in the making of this thesis. After going through the necessary preprocessing we do some basic analysis to get an idea of what exactly is in the data.

3.1 Description of the data

The mobile phone data used in the study was obtained via Notre Dame University in Indiana, USA. It is composed of billing information of a single mobile phone operator in an undisclosed European country. The data consists of two parts, the events data and the aggregated network data, and a separate file with customer demographics.

The events data contains detailed information on all phone calls and SMS messages during January 2007. For each call and SMS, the data contains the IDs of the caller and the callee, exact time of call with an accuracy of one second and the duration of the call.

The aggregated network has been constructed from all phone calls and SMS messages made during the first 18 weeks of 2007, from January

1st to May 6th. Each node in the network represents one customer, and there is a directed edge from customer i to customer j if i has called j at least once or if i has sent j at least one SMS message. The edges are weighted, and there are three different weights: the total number of calls, the total duration of calls or the total number of SMS messages. Note that the aggregated network is a static structure: it has no information about the times the calls took place.

The demographic information contains the following facts about each user:

Age in years

Gender, either male, female or unknown.

ZIP code of the customer

Connection type, either *prepaid* or *postpaid*. The fundamental difference between the two is billing. Prepaid users pay their calls in advance, while postpaid users have made a contract with the mobile operator and are billed regularly on past calls.

The demographic information has much more omissions and errors than the other parts of the data, especially for prepaid users. See Section 3.4 for more information.

Note especially that the data ***does not*** have any information about the contents of calls or SMS messages, nor the true phone numbers or identities of the customers. Individuals are identified only by entirely artificial user IDs. It is practically impossible to link users to real people.

3.2 Preprocessing

3.2.1 Forcing reciprocity

Following the example of [12], we limit our study to the most trusted communication links by removing all edges that only go one way: an edge exists in our final network only if there has been some communication, either calls or SMS messages, in both directions. The aim of this step is to get a better representation of the actual social network by removing all random odd edges. The resulting aggregated network and events data will be referred to as *the reciprocal data*. Because reciprocity is judged according to the aggregated network, the events data is not entirely reciprocal since it spans a shorter time interval (4 weeks instead of 18 in the aggregated network). All discussion in this thesis concerns the reciprocal data unless otherwise noted.

3.2.2 Handling SMS messages

The maximum length of an SMS message is limited to 160 characters by the SMS specifications. Most handsets however allow sending longer messages by dividing them into several parts. From the viewpoint of the operator these parts are all individual messages, and because they are billed as such, they also appear as individual messages in the data. This would cause severe problems if we were to study for instance the lengths of time-intervals between SMS messages. It is more practical to treat the multipart messages as a single messages, which is what the sender intended and what the recipient saw.

Unfortunately the data contains no information whether some set of consecutive SMS messages were in fact the separate parts of one multipart message; we need to make use of the sending times to infer this.

Due to technical reasons SMS messages are not relied instantly. Figure 3.1(a) shows the time interval distribution of the parts of potential multipart SMS messages. Based on this distribution we decide on a time window of 10 sec-

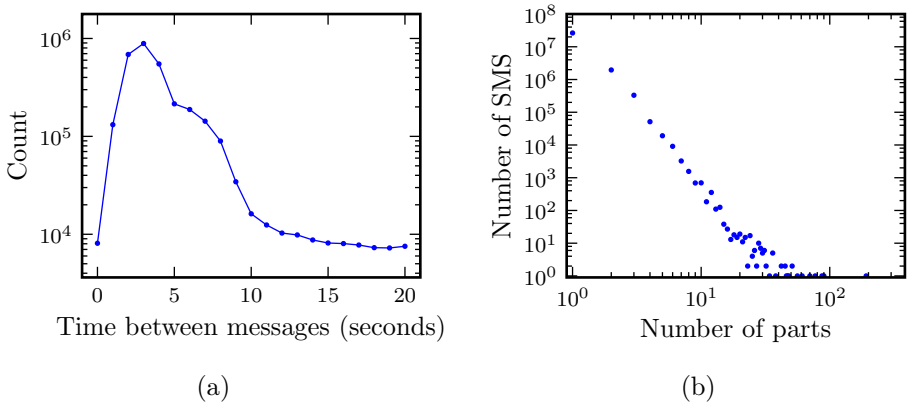


Figure 3.1: **(a)** The distribution of times between individual consecutive SMS messages sent by one user to the same recipient and on the condition that there is no other communication between the messages. When the time interval is very small, it is very likely that the corresponding SMS messages are the parts of one multipart message. **(b)** The distribution of the number of parts in multipart SMS messages when a time window of 10 seconds is used with the above criteria.

onds to judge whether two messages are part of the same multipart SMS message: a series of SMS messages with a time difference less or equal to 10 seconds between two consecutive messages and with no other communication during the whole succession is considered one multipart message. Note that even if this procedure does find some false positives (separate SMS messages that are mistaken for a multipart message), these messages are quite probably either accidentally sent identical messages or quick additions to the previous message, and in both cases we would still like to infer them as a single message in any analysis about information mediation.

Using the criterion above, the parts of a multipart message are replaced by only one SMS message with sending time corresponding to the time of the first part and the *duration* matching the time difference between the first and the last part. This duration is only useful for determining the approximate time taken between sending and receiving an SMS message, and it should not be confused with the duration of a call, which tells the actual time the two people discussed.

Figure 3.1(b) shows the number of multipart messages as a function of the

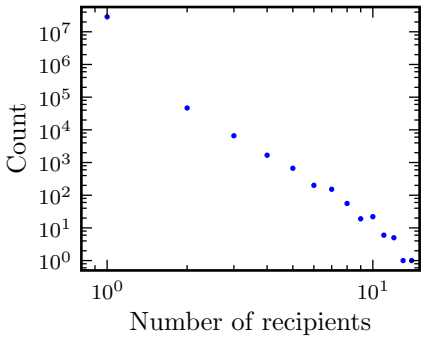


Figure 3.2: The number of recipients of SMS messages. Messages consisting of multiple parts are counted as one.

number of parts. The number of multipart messages is quite significant: within the 31.7 million individual SMS messages billed there are 28.7 million distinct messages. Curiously, the largest multipart SMS has 192 parts. This is probably explained by a binary data file transmitted as several SMS messages, a feature allowed by some advanced handsets.

Most handsets also allow sending text messages to multiple recipients. The data again contains no information whether a user sends the same SMS to multiple recipients or different SMS to many recipients during a very short time period, so we use the same criteria as above and consider the SMS messages identical if they have been sent to multiple recipients within 10 seconds. Figure 3.2 shows the distribution of the number of recipients for SMS messages. About 99.8 % of all messages have been sent to only one recipient, and as a first approximation it is feasible to say that SMS messages are one-to-one communication.

3.3 Basic analysis

After preprocessing the reciprocal data contains a total of 5 343 749 nodes, with 3 227 081 postpaid and 2 054 190 prepaid customers. 62 478 customers with no user type are mostly *churners*, customers who have left the company

during the observation period. In these cases a new customer might have started using the same phone number, but since these cases make up only 1.1 % of the user base, any error caused should be minimal when averaging over large amounts of data.

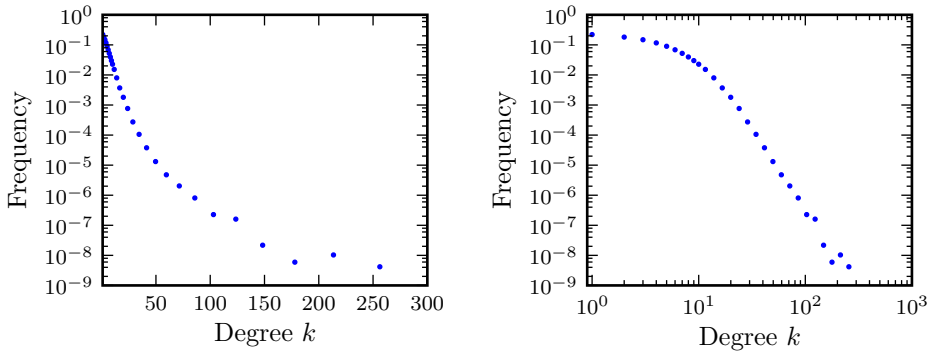
3.3.1 Motivation of the analysis

If the networks studied were very small, our analysis would consist of plotting the network as in Figure 2.1 and describing the network verbally — person A has very few friends, two friends of person B are quite likely to know each other, and so on. While this approach is easy and intuitive, the conclusions would only hold for the network under inspection, and it give us no information about whether the conclusions drawn can be generalized to a larger population. From the complex system point of view, we also ignore all phenomena taking place on larger scales.

To be able to make more accurate and universal conclusion we need much more data. Unfortunately, with millions of people it is no longer possible to just look at the data. Even if we could make a print large enough to accommodate all 5,3 million nodes, we'd have to be quite violent in projecting the network to two dimensions.

The only way to construct a mental image of what large networks look like is to study different properties one or two at a time. The goal of this process is to create an understanding of the structure and dynamics of the network, without actually seeing the nodes themselves.

We start our analysis with the aggregated network. All results are qualitatively applicable to the events data, because the events data could be used to create a subset of the aggregated network, but the use of the aggregated network adds to the precision of the analysis. The temporal information in the events data will then be used to analyse the daily and weekly variations in the data.



(a) (b)

Figure 3.3: Total degree distribution in **(a)** semi-logarithmic and **(b)** double-logarithmic coordinates.

3.3.2 The aggregated network

The complete reciprocal network consists of 350 million phone calls with a total duration of 14.46 million hours and 127 million SMS messages.

Average degree of the network is $\overline{k} = 4.476$. Speaking of the average degree alone is however somewhat misleading. Figures 3.3(a) and 3.3(b) show the total degree distribution in semilogy and loglog coordinates, respectively. It is immediately evident that a large proportion of nodes have a degree that differs very significantly from the average, which is very typical in social networks. We can also see that in Figure 3.3(a) the points lie on a straight line up to degree 20, corresponding to an exponential distribution $p(k) \sim e^{-0.25k}$. From degree 20 onwards the points lie on a straight line in Figure 3.3(b), which corresponds to a power law $p(k) \sim k^{-5.5}$. What these figures do not tell so clearly is that only 0.6 % of the nodes have a degree above 20. The majority of the node degrees are thus exponentially distributed.

Weight and strength distributions

Since we have three possible weights (call count, call length and SMS message count), we can calculate three separate weight and strength distributions,

shown in Figures 3.4(a) and 3.4(b). As the degree distribution, the weight and strength distributions are also fat-tailed.

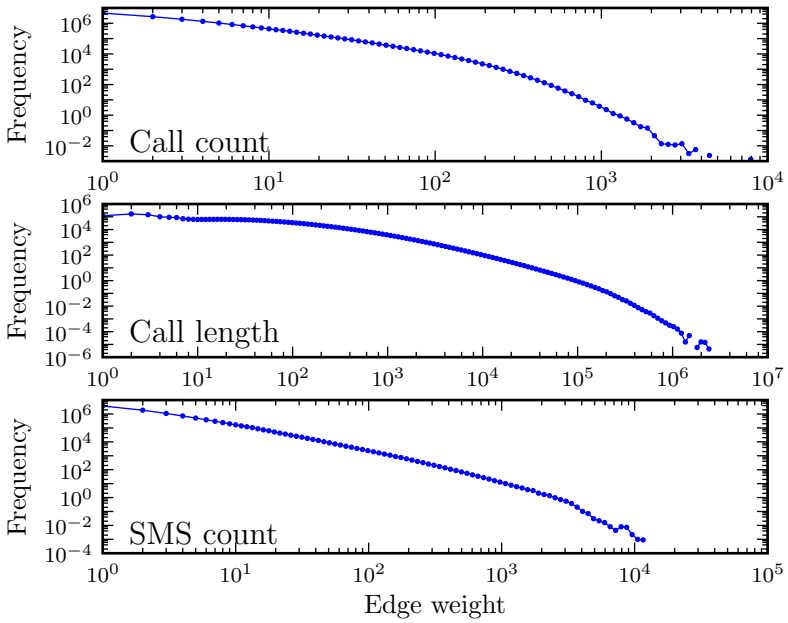
The average strengths are 65.6, 9741, and 23.8 for call count, call length and SMS count, respectively, but it should again be noted that the average of a fat-tailed distribution should not be interpreted as the 'typical value' as is done with the average of a Gaussian distribution. For example with call counts, 69 % of all users make less than the average number of calls, and 0.29 % percent of users make more than 10 times the average number of calls.

3.3.3 The events data

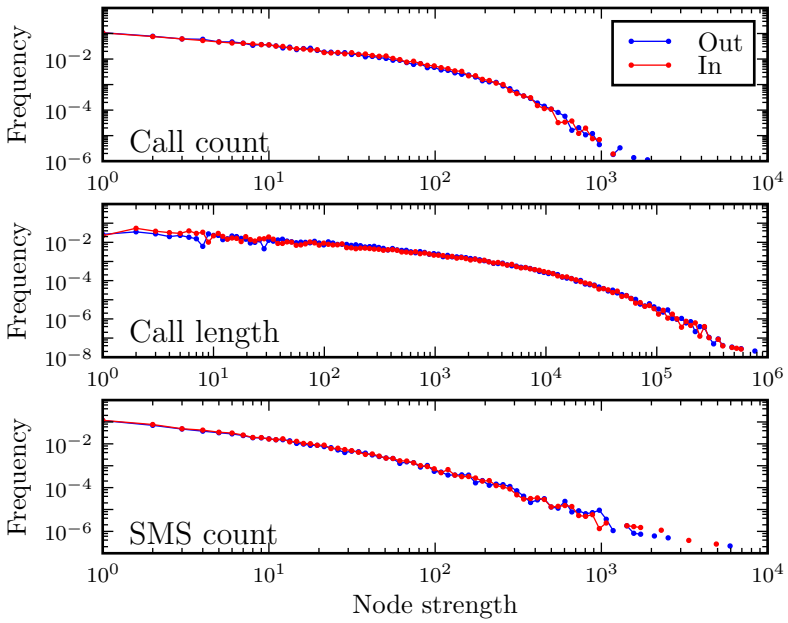
The events data has a total of 83.8 million phone calls and 31.7 million SMS messages, which equals an average of 15.7 calls and 5.93 messages per person during the whole period of 31 days. In the following we will take a closer look at the distribution of calls and SMS messages.

Daily call counts

Quite naturally there are large temporal variations in average calling behaviour. Figure 3.5(a) shows the number of calls and Figure 3.5(b) the number of SMS messages on each day. The January 1st is the New Year in all European countries, which can be seen as a pronounced spike in the number of SMS sent, but curiously not in the number of calls. The calls have a very clear weekly pattern, with Fridays and Saturdays having the most calls and Sundays being the most quiet. With SMS messages Saturdays and Sundays have the smallest number of traffic, which is a first hint about the different role of calls and SMS messages in mobile phone communication.

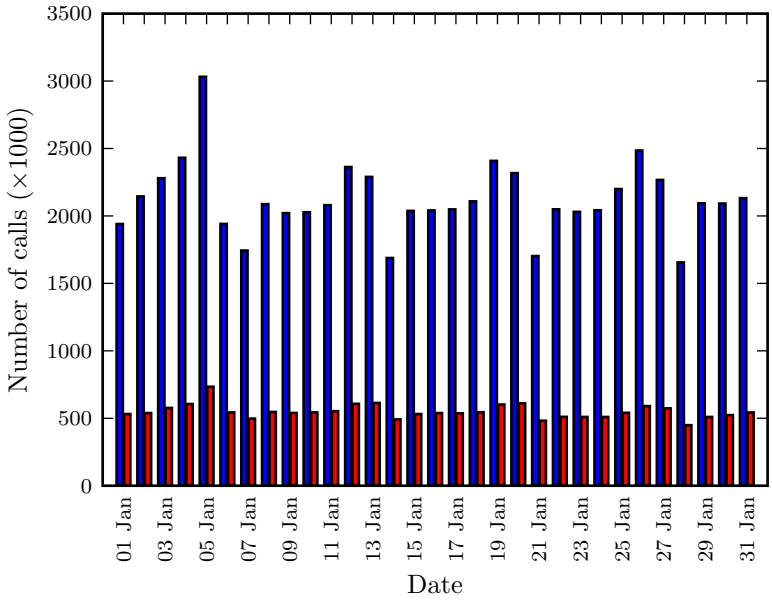


(a)

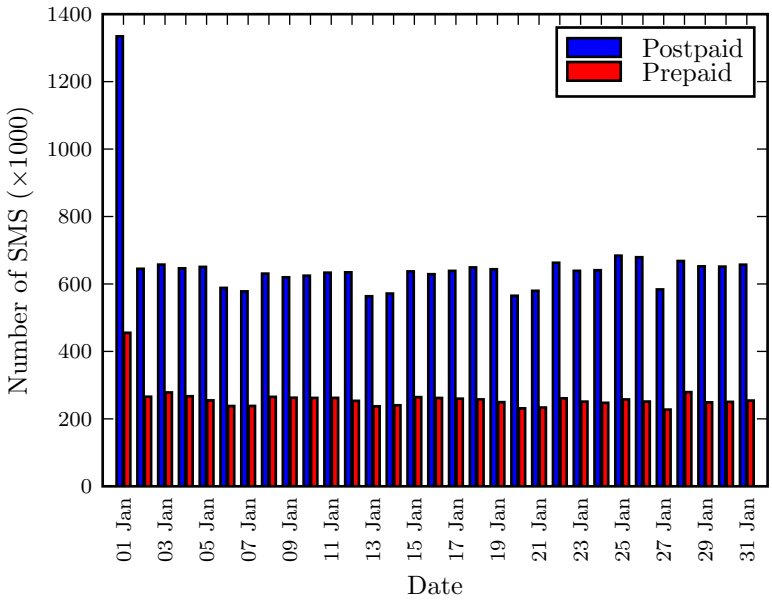


(b)

Figure 3.4: **(a)** Weight distributions for the weights of directed edges for all three weight types. **(b)** In- and out-strength distributions for all weight types.



(a)



(b)

Figure 3.5: Daily statistics in January 2007 of **(a)** the total number of phone calls and **(b)** the total number of SMS messages, shown separately for prepaid and postpaid users. Monday January 1st is the New Year, which is seen a spike in the number of SMS messages sent. Multipart messages are counted as only one, but messages to multiple recipients are counted according to the number of recipients.

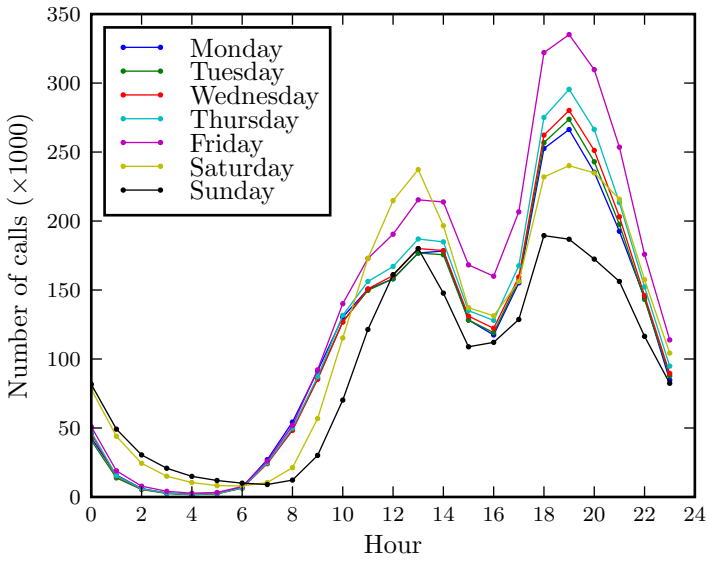


Figure 3.6: Average call counts by hour for each weekday. January 1st is excluded from the data to remove the effect of calls on New Year. Note that the graphs continue to the next weekday, for example the next point after the last hour on Monday is the first point on Tuesday.

Hourly patterns

Figures 3.6, 3.7 and 3.8 show the average hourly variation of call counts, call lengths and SMS message counts, respectively.

Figure 3.6 shows clearly how weekdays differ from the weekend. Weekdays from Monday to Thursday are quite similar, with the number of calls increasing as the week advances. People call more during office hours and in the evening. Friday has a lot more activity, probably in anticipation of the weekend, and looking at the number of calls after midnight we can see that the daily cycle moves forward by about two hours during the weekend. Sunday afternoons are the most quiet.

The average call lengths in Figure 3.7 also reveal the difference between the work week and the weekend. The largest difference is however between day and night. The average call during the night on weekdays is about 4 times

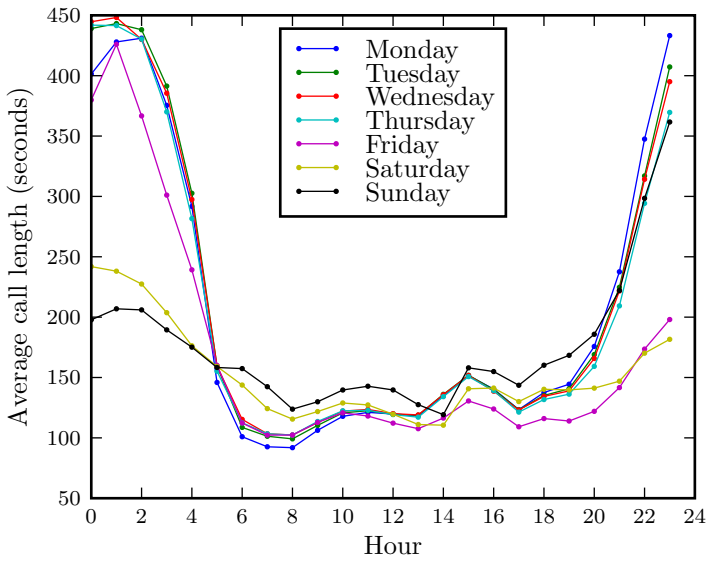


Figure 3.7: Average call lengths for each weekday. January 1st is excluded from the data to remove the effect of calls on New Year.

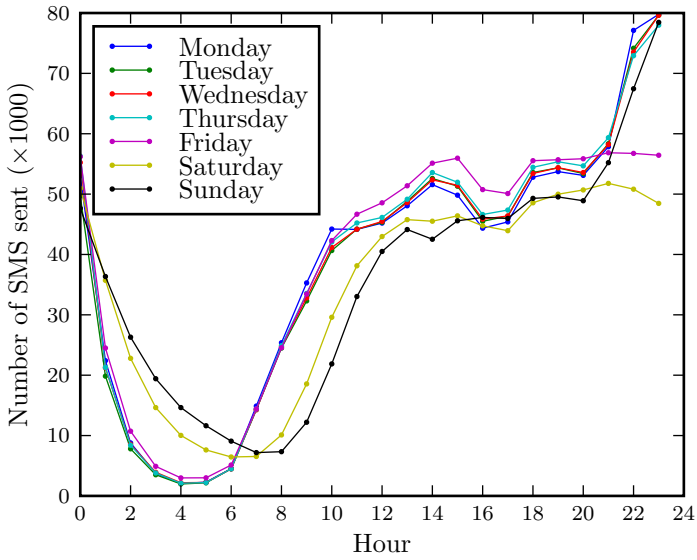


Figure 3.8: Average SMS message counts by hour for each weekday. January 1st is excluded from the data to remove the effect of calls on New Year.

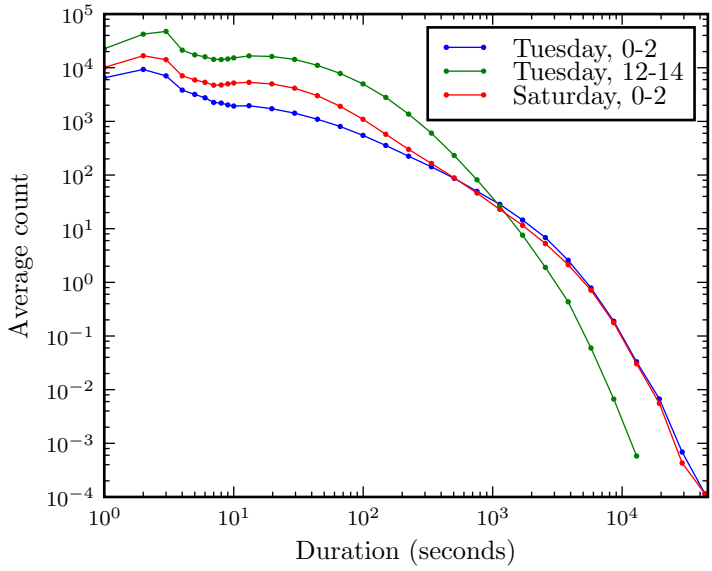


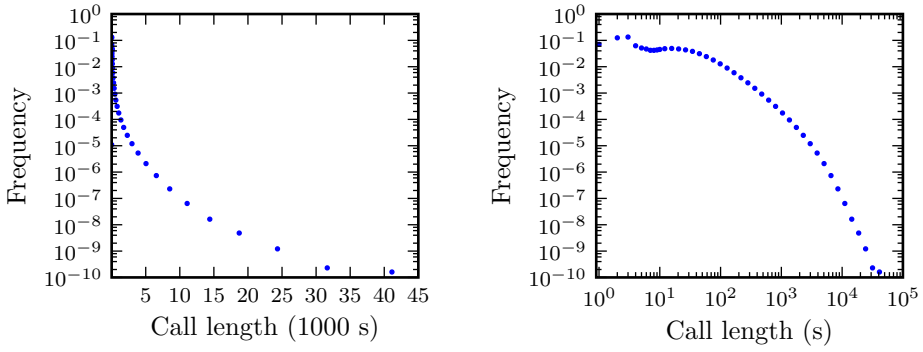
Figure 3.9: The number of calls of different durations on three time intervals.

as long as during the day. This curious feature necessitates a closer study.

Figure 3.9 shows the number of calls of different durations during three different time intervals. Comparing the plots for Tuesday 0:00-2:00 and Tuesday 12:00-14:00 we can see that the longer average call durations during night stem from two differences: the number of short calls is reduced and the number of long calls is increased, and the longest calls are also longer than during the day. Comparing the plots for Tuesday 0:00-2:00 and Saturday 0:00-2:00 shows that the number of long calls is exactly the same on both days; the shorter average call duration during the weekend is explained by the larger number of short calls.

Call length distribution

Figures 3.10(a) and 3.10(b) show the total call length distribution in semi- and double logarithmic coordinates, respectively. Unlike the degree distribution, these plots do not show any straight lines; the functional form of the call length distribution is somewhere between an exponential and a power



(a) (b)

Figure 3.10: Call length distribution in **(a)** semi-logarithmic and **(b)** double-logarithmic coordinates.

law. Yet the distribution are clearly fat-tailed, as the distribution spans over several orders of magnitude.

3.4 Problematic features

After taking a look at what *is* included in the data it is appropriate to discuss what is *not*.

Most constraints for the usability of the data stem from the fact that the data was not originally collected for research purposes, but to offer sufficient information for billing; the company understandably has little interest to spend money on collecting data about its customers as long as they pay their bills in time.

The demographic information has many missing values. Since getting a prepaid subscription is as simple as buying a ready-made package from a store, the prepaid users have little incentive to give their personal information to the phone company. Only about 40 % of prepaid users have a valid age, gender and zip code in the data, while nearly 99 % of postpaid users have full information. Table 3.4 shows the number of users according to user type and gender. The large number of male prepaid users is quite likely due to the fact

Table 3.1: The number of users according to user type and gender. All numbers in thousands of people.

		Gender			
		Unknown	Male	Female	**Total**
Type	Postpaid	2	1 394	1 831	3 227
	Prepaid	6	390	1 658	2 054
	Missing	0.2	20	42	62
	Total	8	1 805	3 531	**5 344**

that it is the default gender for those who have not specified one.

Under-aged users are unable to obtain postpaid subscription in person, and it seems that their phones are often listed under their parent's name. This can be seen for example in Figure 3.11, which has an unnaturally strong diagonal. This doesn't affect only calls between parents and children: any call between two teenagers would be logged as a call between their parents.

Because the data consists of billing data, it has no information of unanswered calls. Also since the users are identified by the phone number, the actual person using this number might change. During the 18 week period, there are a total of 70687 such changes, which luckily makes the effect small enough to be averaged out in most analysis.

All these defects of the data place limits to what we can and should do with the data set. To make sound conclusions, we should concentrate on questions and answers that make use of the more accurate part of the data: the precise and complete information of calls and SMS, including their exact time and duration.

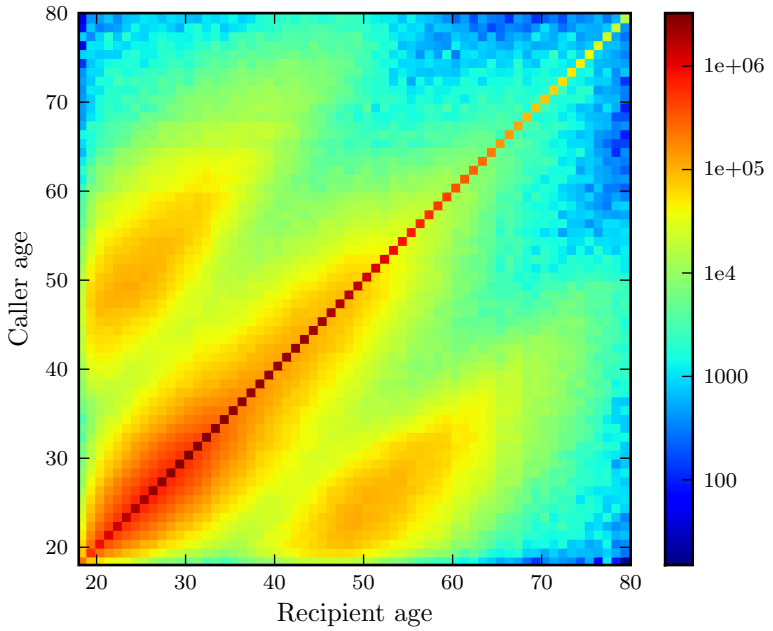


Figure 3.11: The total number of calls between age groups for postpaid users. The strong diagonal is an artifact, most likely caused by calls inside families where the contracts of the children contain the demographic information of their parents. The two clusters around the diagonal correspond to calls between parents and their children (in cases where the demographic information is correct).

Chapter 4

Reciprocity of edges

One rather little studied aspect of directed networks is link reciprocity: how often do directed edges go both ways between two nodes. This is an interesting question with respect to social networks because reciprocity can be seen as a measure for the evenness of relationships.

For unweighted directed networks the reciprocity is commonly defined as the fraction of edges that point both ways [23]. It is known that nearly all unweighted social networks have a high degree of reciprocity, generally several orders of magnitude higher than if the edges were completely random, and this result is sometimes used to transform weighted, directed networks into undirected networks by simply taking the average weight of each 2-way edge. However, what has hitherto been very little studied is edge reciprocity in *weighted* directed networks.

4.1 The edge bias

With a weighted network we can study reciprocity more closely than in the unweighted case. A very simple measure (and probably the most intuitive one) for the reciprocity of weighted edges is the fraction of total weight on

one edge,

$$b_{ij} = \frac{w_{ij}}{w_{ij} + w_{ji}} \quad . \tag{4.1}$$

We will call the quantity b_{ij} the *edge bias* of edge (i, j). Obviously $b_{ij} = 0.5$ when $w_{ij} = w_{ji}$, $b_{ij} = 1$ when $w_{ji} = 0$, and $b_{ij} + b_{ji} = 1$.

The first interesting question is the form of the edge bias distribution. Is it a reasonable hypothesis that the total weight is equally distributed on both edges? If not, what causes the uneven weight pattern? Are there regularities and correlations that explain the distribution of biases? These are the question we try to answer in this chapter.

This chapter will only deal with reciprocity with respect to call counts. The number of SMS messages could be used similarly, but because there are differences in the usage patterns of the two media, the results would probably not be identical.

Note that when studying the reciprocity there is a remarkable difference between the number of calls and the total duration of calls. The number of calls is a measure of activity: if during the 18 week period A calls B 100 times but B calls A only 50 times, it is quite natural to think that A is more active (takes the initiative more often) in this relation. However, if the total duration of the calls made by A is twice that of those made by B, we still do not know who did the talking; it could well be that both were speaking for the same amount of time. To measure reciprocity with call duration we'd need to know how much each person was speaking, and this piece of information we do not have.

4.1.1 Variation of edge bias

Figure 4.1(a) shows the distribution of edge bias values as a function of total edge weight. Because $b_{ij} = 1 - b_{ji}$, the distribution is be symmetric around $b_{ij} = 0.5$, and therefore it suffices to study only the values $b_{ij} \geq 0.5$. Note that the edge bias is quantized in the low end: for example, if the total edge weight

(total call count in this case) is 5, the more active participant can make 3, 4 or 5 calls, corresponding to edge bias values 0.6, 0.8 and 1. Excluding the smallest total weights, the distribution of edge bias does not seem to depend very strongly on the total weight.

To avoid the problem with quantization we limit our study to edges with $50 \leq W_{ij} \leq 1000$, where $W_{ij} = w_{ij} + w_{ji}$, corresponding to the reasonably stable middle part in Figure 4.1(a). Note that while these edges make up only 15.3 % of all edges, they relay 68 % of all calls (see Fig 4.1(b)). Figure 4.1(c) shows the cumulative edge bias distribution for these edges. The distribution has two nearly linear segments, the first one in the range $0.5 \leq b_{ij} \leq 0.65$ (approximately 50 % of the probability mass) and the second one in the range $0.8 \leq b_{ij} \leq 1.0$ (20 % of the edges). Linearity of the cumulative distribution means that the probability density is approximately uniform in these ranges. The observed uniform distribution is already a large deviation from the hypothesis that the calls were evenly distributed on both edges.

Of course it is not very clever to expect that *all* edges had exactly the same number of calls in both directions. A more realistic claim is that both people have the same probability of making a call, which would mean that the number of calls in either direction follows a binomial distribution with parameters $n = W_{ij}$ and $p = 0.5$. If this hypothesis were true, we should be able to find several edges with a large bias when W_{ij} is small, but the probability of observing a large bias should decrease exponentially with W_{ij}.

Figure 4.1(d) shows what the edge bias plot would look like with the binomial hypothesis. Obviously this is still far from the observed distribution in Fig 4.1(a). When $W_{ij} \geq 100$, edge bias values of over 0.75 are practically nonexistent in the binomial case while in the actual data such edges are plentiful.

4.1.2 Edge bias and the strength distribution

Assuming the edges to be even in the statistical sense turned out to be too strong an assumption. But given the network topology and the fat-tailed

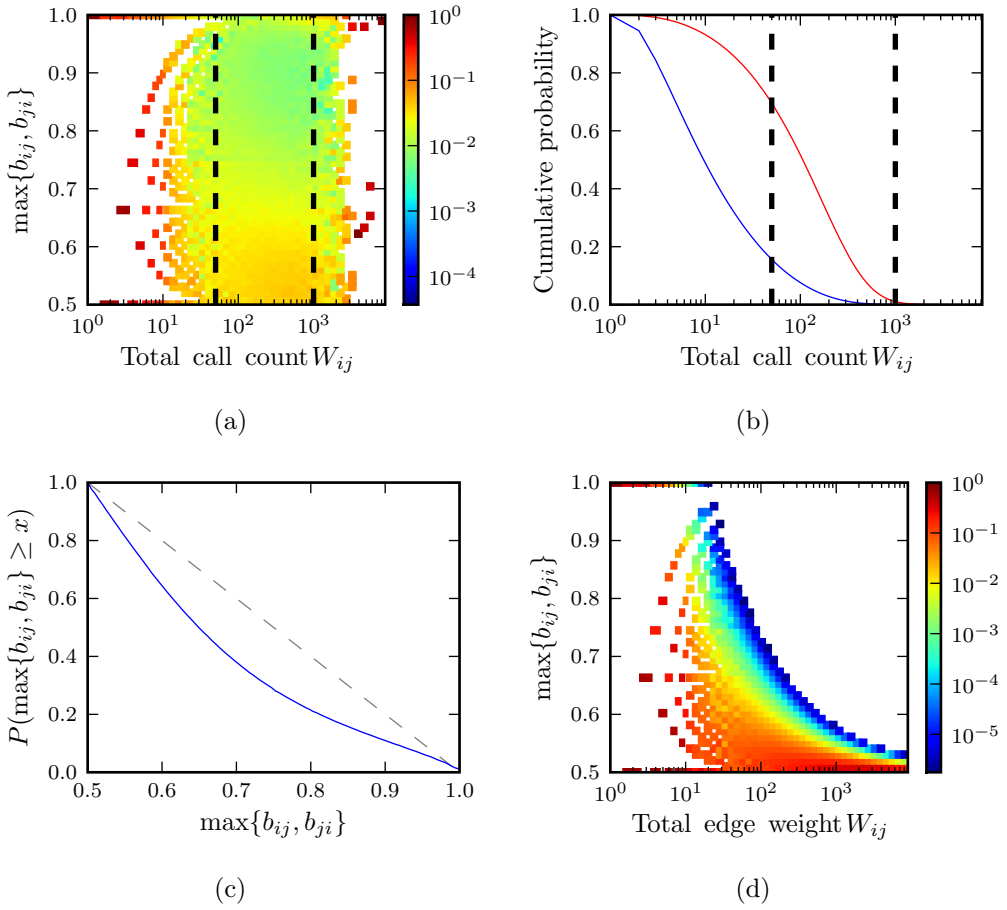


Figure 4.1: **(a)** The distribution of edge biases as a function of edge weight. Each column sums to 1 and shows the distribution of biases of edges with similar total weights; in other words, each column shows the distribution $p(\max(b_{ij}, b_{ji})|\underline{w} \leq W_{ij} < \overline{w})$. The black dashed lines illustrate the range $50 \leq W_{ij} \leq 1000$ that is analysed more closely in (c). The columns in this range are very similar, which means that the distribution is roughly independent of W_{ij}. **(b)** The cumulative distribution for the fraction of edges (blue) and the fraction of calls (red) as a function of total weight W_{ij}. The black dashed lines again illustrate the range $50 \leq W_{ij} \leq 1000$, and we see that while the edges in this range make up only 15.3 % of all edges, those edges relay 68 % of all calls. **(c)** The cumulative distribution of call count bias for all edges with $50 \leq W_{ij} \leq 1000$ (blue). The diagonal marks the cumulative distribution of a uniform distribution. **(d)** The distribution of edge bias for call counts if both people in each relationship had an equal chance of making a call.

strength distribution, is it possible for the bias distribution to be even? If we use the strength of a node as a measure of activity, wouldn't the large variations in personal activity be enough to explain the bias distribution? If this were the case, the bias would no longer be a true property of an edge, but a simple consequence of the heterogeneity of nodes.

To see whether the strength distribution alone is enough to cause the bias distribution we try to redistribute the total out-strength of each node on its outgoing edges so that the resulting edge biases are as even (close to 0.5) as possible. If it turns out that such redistribution is not possible, we may conclude that the strength distribution is a sufficient explanation for the observed biases.

Problem definition

To even out the biases throughout the network we select to maximize the likelihood that the biases come from a binomial distribution with $p = 0.5$. This naturally forces the observation that a large bias is quite likely to occur in an edge with a small weight but very unlikely in an edge with a large weight. Thus we attempt to find the a new set of weights $\mathbf{w} = \{w_{ij}\}_{(i,j)\in\mathbf{E}}$ such that

$$\begin{aligned}\mathbf{w} = \arg\max \Pi_{(i,j)\in\mathbf{E}} &\Pr(w_{ij} |\, w_{ij} \sim \mathrm{Bin}(w_{ij} + w_{ji},\, 0.5)) \\ \text{s.t.} \sum_j w_{ij} &= s_i \quad .\end{aligned}$$

The topology of the original network must be retained, meaning that weight may be moved from edge $(i,\, j)$ to edge $(i,\, k)$ only if $w_{ik} > 0$ in the original network. (We'll explain later why it is clever to maximize a global function instead of just making some local changes to even out the bias. Just stick with it for a while.) What follows is a description of the method used to solve the problem. The complete explanation is a bit heavy on equations, and less mathematically-oriented readers may skip straight to the results.

By writing the binomial probability explicitly, we get

$$\begin{aligned}
\mathbf{w} &= \arg\max \sum_{(i,j)\in\mathbf{E}} \log \Pr(w_{ij} | w_{ij} \sim \mathrm{Bin}(w_{ij} + w_{ji},\, 0.5)) \\
&= \arg\max \sum_{(i,j)\in\mathbf{E}} \log \frac{W_{ij}!}{w_{ij}! w_{ji}!} p^{w_{ij}} (1-p)^{w_{ji}} \\
&= \arg\max \sum_{(i,j)\in\mathbf{E}} \left(\sum_{u=1}^{W_{ij}} \log u - \sum_{u=1}^{w_{ij}} \log u - \sum_{u=1}^{w_{ji}} \log u + w_{ij} \log p + w_{ij} \log(1-p) \right) \\
&= \arg\max \sum_{(i,j)\in\mathbf{E}} \left(\sum_{u=1}^{W_{ij}} \log u - \sum_{u=1}^{w_{ij}} \log u - \sum_{u=1}^{w_{ji}} \log u \right) \quad .
\end{aligned}$$

The last equality follows from the fact that with $\log p = \log(1-p)$ since $p = 0.5$ and $\sum_{(i,j)\in\mathbf{E}} w_{ij} + w_{ji}$, stays constant when the weights are redistributed. We may now define the global target function

$$f(\mathbf{w}) = \sum_{(i,j)\in\mathbf{E}} \left(\sum_{u=1}^{W_{ij}} \log u - \sum_{u=1}^{w_{ij}} \log u - \sum_{u=1}^{w_{ji}} \log u \right) \quad . \tag{4.2}$$

Solving the problem

The global target itself is not enough to solve the problem — we must still find a way to maximize it. This is an optimization problem of the worst kind, with millions of integer variables and a myriad of constraints, and finding a global optimum could require going through all valid weight combinations. However, by exploiting the constraint that the strength of each node must be preserved we can create a simple algorithm that improves the value of the target function one step at a time.

The simplest change of weights that preserves the strength of node i is to move one unit of weight from edge $(i,\, j)$ to $(i,\, k)$. The value of $f(\mathbf{w})$ increases

if $\Delta f(\mathbf{w})$ is positive:

$$\begin{aligned}
\Delta f(\mathbf{w}) &= f(\mathbf{w};\, w_{ij} \mapsto w_{ij} - 1,\, w_{ik} \mapsto w_{ik} + 1) - f(\mathbf{w}) \\
&= \sum_{u=1}^{w_{ij}+w_{ji}-1} \log u - \sum_{u=1}^{w_{ij}+w_{ji}} \log u + \sum_{u=1}^{w_{ij}} \log u - \sum_{u=1}^{w_{ij}-1} \log u \\
&\quad + \sum_{u=1}^{w_{ik}+w_{ki}+1} \log u - \sum_{u=1}^{w_{ik}+w_{ki}} \log u + \sum_{u=1}^{w_{ik}} \log u - \sum_{u=1}^{w_{ij}+1} \log u \\
&= -\log(w_{ij} + w_{ji}) + \log w_{ij} + \log(w_{ik} + w_{ki} + 1) - \log(w_{ik} + 1) \\
&= \log \frac{w_{ij}(w_{ik} + w_{ki} + 1)}{(w_{ij} + w_{ji})(w_{ik} + 1)} = \log \frac{b_{ij}(0)}{b_{ik}(1)} > 0 \quad ,
\end{aligned}$$

where

$$b_{ij}(r) = \frac{w_{ij} + r}{w_{ij} + w_{ji} + r}$$

is the bias of the edge $(i,\, j)$ after adding weight r. This final result is surprisingly simple: the target function value is increased whenever $b_{ij}(0) > b_{ik}(1)$. Note that while $f(\mathbf{w})$ was derived by maximizing the likelihood of binomial distributions, we end up comparing the edge biases. The two biases compared implicitly account for the form of the binomial distribution; see Figure 4.1.2 for explanation.

There is no reason to limit switching only one unit of weight at a time. Since the first unit is moved if $b_{ij}(0) > b_{ik}(1)$, it takes little thought to realise that another unit of weight should be moved if $b_{ij}(-1) > b_{ik}(2)$. More generally, r units of weights should be moved if $b_{ij}(1-r) > b_{ik}(-r)$; the largest integer r for which this condition is true gives the total amount of weight that should be moved from $(i,\, j)$ to $(i,\, k)$, which can be written as

$$r = \left\lceil \frac{w_{ij}w_{ki} - w_{ji}w_{ik} - w_{ji}}{w_{ji} + w_{ki}} \right\rceil \quad . \tag{4.3}$$

It should now be obvious why it's very beneficial to maximize a global target function instead of trying to even out the bias locally. Because

1. there is a finite number of possible ways to distribute the discrete

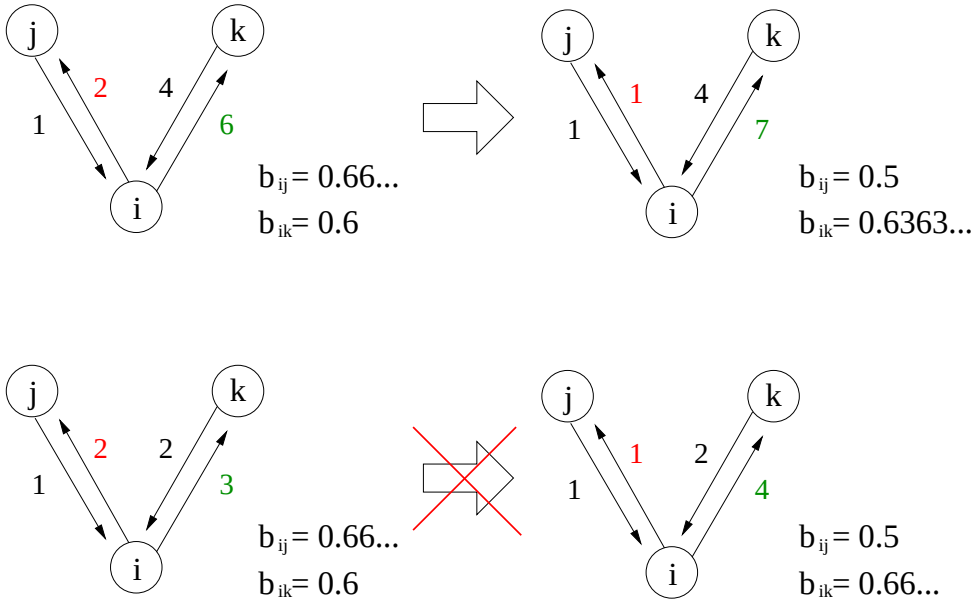


Figure 4.2: **(Top)** One unit of weight is moved from edge (i, j) to (i, k). Since $b_{ij}(0) = 0.6\overline{6} > 0.\overline{63} = b_{ik}(1)$, this change increases the value of the target function. This means that the relative increase in the likelihood of the edge (i, j) (33.3 %, from $\binom{3}{2}0.5^3 = 0.375$ to $\binom{2}{1}0.5^2 = 0.5$) outweighs the decrease in the likelihood of the edge (i, k) (21.4 %, from $\binom{10}{6}0.5^{10} \approx 0.205$ to $\binom{11}{7}0.5^{11} \approx 0.161$): $(1 + 0.3\overline{3})(1 - 0.214) > 1$. **(Bottom)** Because $b_{ij}(0) = 0.6\overline{6} = b_{ik}(1)$, this move does not increase the value of the target function and is therefore not performed. In fact, since the two bias values turn out to be equal, this means that the increase in the likelihood of the edge (i, j) (33.3 %) exactly matches the decrease in the likelihood of the edge (i, k) (25 %, from $\binom{5}{3}0.5^5 = 0.3125$ to $\binom{6}{4}0.5^6 \approx 0.234$): $(1 + 0.3\overline{3})(1 - 0.25) = 1$

weights while retaining the network topology and the strength of each node, and

2. the value of $f(\mathbf{w})$ is increases at each step

it follows that there can not be cycles and we are guaranteed to find a local optimum after a finite number of steps.

We can now outline an algorithm that evens out the bias weights. We start with an initial network which defines the network topology and the strength of each node. Instead of starting with the original network it might be useful

to redistribute the out-weights in relation to the in-weights. The algorithm consist of several rounds. On every round we then go through each node i, and select the neighbours $j^* = \arg_j \max\{b_{ij}(0)\}$ and $k^* = \arg_k \min\{b_{ik}(1) | k \neq j^*\}$. If $b_{ij^*}(0) \leq b_{ik^*}(1)$, we cannot improve the current node and we proceed to the next node.[1] Otherwise we move weight r defined by equation (4.3) from edge (i, j) to (i, k), reselect nodes j^* and k^* according to the new weights and see if we can still improve further.

The algorithm finishes when no change of weight is made during one full round. Note also that on each round it is only necessary to check the nodes whose incoming edge weights have been changed during the previous round. In addition we can skip all nodes with degree less than 2 since in this case it is not possible to switch weights.

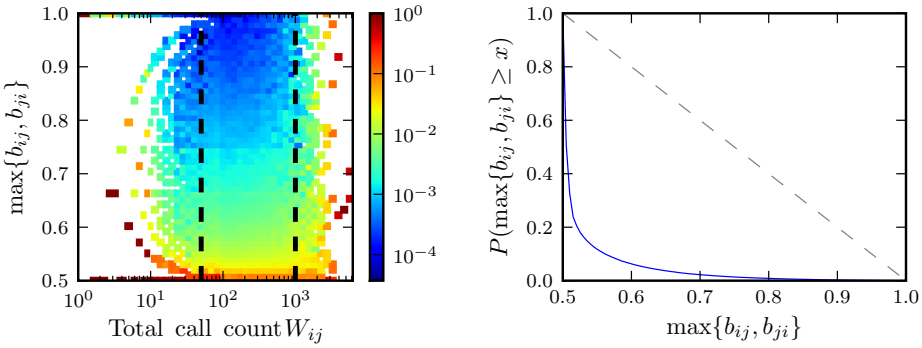
The results

The bias distribution for the new network is shown in Figure 4.3(a) and the cumulative distribution of biases of edges with total weight between 50 and 1000 is shown in Figure 4.3(b). The bias values are now heavily concentrated around 0.5. While in the original network about 65 % of the edges had a bias of over 0.6, in the new network such edges make up only about 7 % of all edges.

The result is quite obvious. Comparing Figures 4.1(c) and 4.3(b) we can see that it's quite possible to significantly reduce the amount of strongly biased edges while retaining the original strength distribution. Therefore even though the strength distribution undoubtedly does contribute to the existence of large bias values, it is in no way a sufficient explanation.

Note that even though the solution reached may only be a local optimum, it is not necessary to find a stronger solution. The result shows that the strength distribution alone does not force the large edge bias values; a better

[1] Note that if $b_{ij^*}(0) \leq b_{ik^*}(1)$ then $b_{ij}(0) \leq b_{ik}(1) \quad \forall j \neq k$. This follows from the definition of j^* and k^*.



(a) (b)

Figure 4.3: **(a)** The edge bias distribution after the weights have been evened out. **(b)** The cumulative distribution of call count bias for edges in the approximately constant part of the bias distribution ($50 \leq W_{ij} \leq 1000$) which now make up 30.3 % of all edges and relay 79.1 % of all calls.

optimum could only further confirm this result.

4.1.3 Significance of end degrees

It would appear that the abundance of large edge biases is not a simple consequence of node properties but a true feature of the edges themselves. The natural follow-up question is whether we can identify regularities in the bias values with respect to other local properties of the network.

One possible source of correlation is the degrees of the end nodes. Figure 4.4(a) shows the average edge bias as the function of the degree of adjacent nodes,

$$\overline{b}_{k_c,\,k_r} = \frac{1}{N_{k_c,\,k_r}} \sum_{\substack{(i,\,j)\in\mathbf{E} \\ k_i=k_c,\,k_j=k_r}} b_{ij}\,, \qquad N_{k_c,\,k_r} = |\{(i,\,j)\in\mathbf{E}\,|\,k_i=k_c,\,k_j=k_r\}| \quad , \tag{4.4}$$

where the k_c and k_r are the degrees of the caller and the recipient, respectively. While we it can already be seen that the edge bias is on average positive when $k_c > k_r$, this is more obvious in Figure 4.4(b), where we show

the *weighted* average bias, defined as

$$\overline{b}^{\mathrm{w}}_{k_c,\,k_r} = \frac{1}{W_{k_c,\,k_r}} \sum_{\substack{(i,\,j)\in\mathbf{E} \\ k_i=k_c,\,k_j=k_r}} (w_{ij} + w_{ji}) b_{ij}\,, \qquad W_{k_c,\,k_r} = \sum_{\substack{(i,\,j)\in\mathbf{E} \\ k_i=k_c,\,k_j=k_r}} w_{ij} + w_{ji} \quad . \quad (4.5)$$

Because the weighted averages are larger than plain averages, we conclude that when $k_c > k_r$, the large weights go hand in hand with large bias values. This conclusion gains more evidence from Figure 4.4(c), where we calculate the average from edges with $10 \leq w_{ij} + w_{ji} \leq 1000$ — removing the edges with small weights further increases the average bias when $k_c > k_r$, which means that the small bias values are more common with small weights. Figure 4.4(d) has the same weight range but with a weighted average.

One should note that the differences in the average bias values are not gigantic; for instance, the average bias between nodes of degree 5 and 10 is 0.484, while the standard deviation of the bias is 0.267 (weighted average is 0.472 with standard deviation 0.228). The differences might seem small, but consider the following:

- The diagonal is equal to 0.5 by definition — if the nodes i and j have equal degree, averaging out b_{ij} and $b_{ji} = 1 - b_{ij}$ gives 0.5. It is however not obvious why the edge bias should change monotonously when the degree of i or j is altered.
- The average edge bias remains above 0.5 for a very large range of end degree values. Nodes with degree ≤ 30 make up almost 99.9 % of all nodes.

4.2 Discussion

In this chapter we have defined *the edge bias*, a new measure for quantifying the reciprocity of directed, weighted edges. The chosen measure is both simple

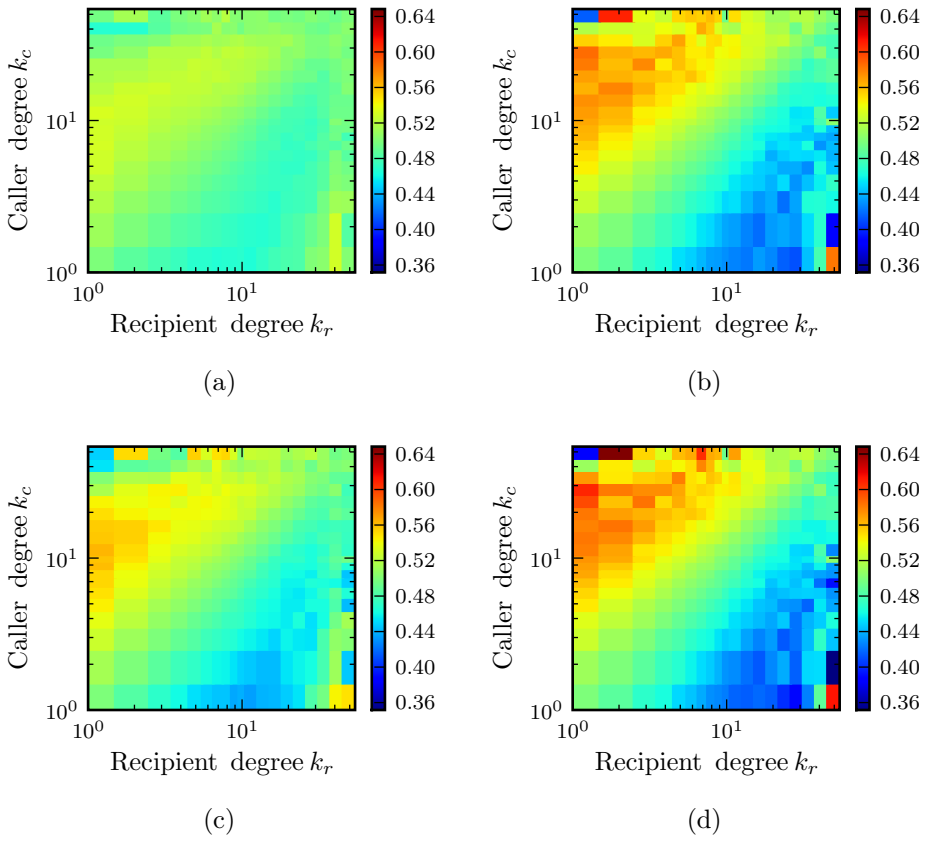


Figure 4.4: **(a)** Average edge bias as a function of caller (vertical axis) and recipient (horizontal axis) degrees. Nodes with degree larger than 50 are not shown because there are only very few such nodes and therefore the data becomes noisy. Note that the plot is anti-symmetric around the diagonal: $\overline{b}_{k_c,k_r} = 1 - \overline{b}_{k_r,k_c}$. **(b)** Weighted average edge bias as defined by Eq. (4.5). **(c)** Average edge bias of edges with total weight between 10 and 1000. **(d)** The same as previous but with weighted average.

and easily interpreted, yet it manages to catch non-trivial properties of the network.

It was first seen that the edge bias values exhibit large variation, and the distribution is quite far from the one we get if we assume that each participant has the same probability to make a call. In many applications the edges are simply assumed to be even — the error made by this assumption of course depends greatly on the nature of the task, but it could be significant especially if the edge bias turns out to have regularities with respect to other network

properties. The directedness of edges would for instance have large effects on the spreading of information in the network.

It was then shown that the fat-tailed strength distribution is not a sufficient explanation for the variance of the edge bias. The effect of the strength distribution might be diminished by the fact the strengths of neighbours are correlated; people who call much know other people who call much.

Lastly we saw that the edges tend to be biased in such a way that the node with a higher degree has a larger contribution to the total weight, with a growing difference in bias as the difference of the degrees grows. This bias difference was also shown to grow with the the total weight of the edge. This shows that there is indeed some regularity in the reciprocity. Another possible source of bias is the mundane user type — postpaid users make on average more calls than prepaid users, which should be reflected in the bias.

This short introduction should establish the edge bias as a useful measure of the reciprocity of weighted, directed edges and also show the rather large lack of reciprocity in mobile communication. It is a question of great interest whether the observed lack of reciprocity extends beyond to communication behaviour studied here. Are human relations inherently biased one way or the other?

Chapter 5

Causality

Because of the high resolution and accuracy of the time stamped data our analysis in not limited to structural properties of the network — we may also study processes taking place in time. In this chapter we'll take a closer look at the causality of mobile communication behaviour. Our aim is to identify to what extent incoming calls (or SMS messages) can be said to *cause* outgoing calls.

The answer to the question will of course be statistical in nature. Looking at the sequence of incoming and outgoing calls of any single individual, it is nearly impossible to say what exactly the made this individual place a call at any given time. To answer the question of causality at this level we'd need to ask the person herself, and even she would probably be unable to specify exactly why the thought of making a call occurred at that precise moment.

5.1 Action triggers

The method we use to study causality is the so called *action trigger* plot, motivated by studies of real neurons as described in [24]. Because measurement of neurons is inherently noisy, the activation stimulus of a neuron is studied

by averaging over several measured stimulus aligned at the exact moment the neuron activates. Even though individual signals are noisy, we can find the effective stimulus by averaging over a large enough sample.

Instead of electric potentials and neurons we study phone calls and the people making them. To create an action trigger plot we align the calling times of all outgoing calls and plot the total number of incoming calls (with an accuracy of one second) before each outgoing call — see Figure 5.1 for a detailed explanation. Thus, if there is *a characteristic time* from the end of an incoming call to the beginning of an outgoing call, we should see it as a spike in out plot.

The characteristic time naturally depends on the type of communication — responding by writing an SMS takes time to compose the message, while a phone call can be made in a matter of seconds. In addition to such technical differences there may be differences in usage habits.

Figures 5.2(a) through 5.2(d) show the action trigger plots for all four possible combinations of incoming and outgoing message types. The top plot in each figure shows returned calls or SMS messages, that is, cases where person A first calls B, and then B calls A. The bottom plot show call or SMS messages to a third party, cases where person A first calls B and then B calls to C.

We can see that almost all plots in Figures 5.2(a)-5.2(d) show the expected spiking behaviour to some extent. For instance, in 5.2(a) the maximum occurs at 17 seconds for returned calls and 25 seconds for calls to a new person. The extra time needed on average to make a call to a new person could be because of technical reasons (some handsets allow fast calling to recent numbers) or because of time taken to mentally prepare for the new call.

The plots show several interesting features. All plots have a minimum at approximately 12 hours (43 200 s), corresponding to natural circadian rhythm — since most calls are made during the day, there are few incoming calls 12 hours before during the night. The lower plot in Figure 5.2(c) also shows curious behaviour. The constant part up to 10 seconds reflects the time it

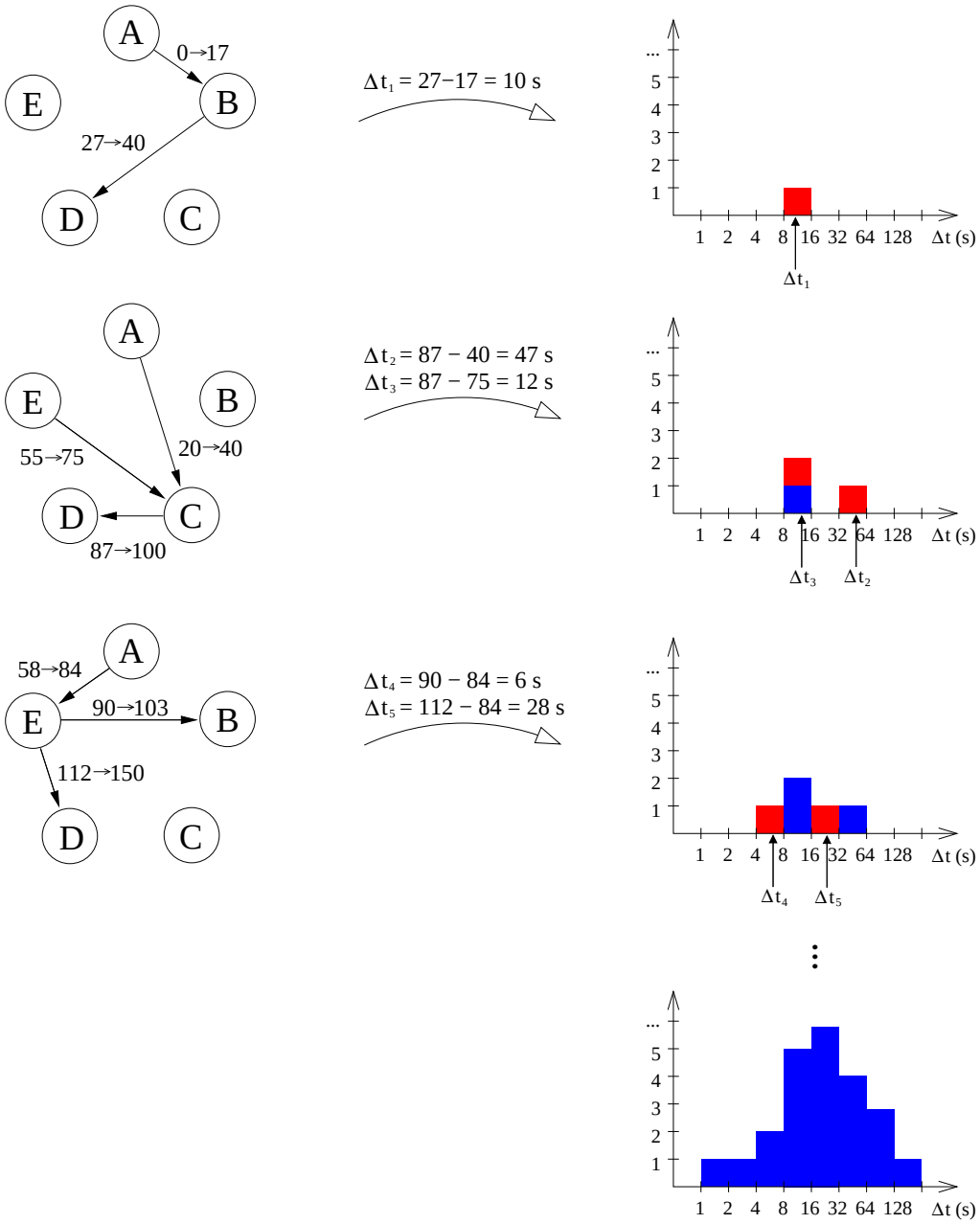
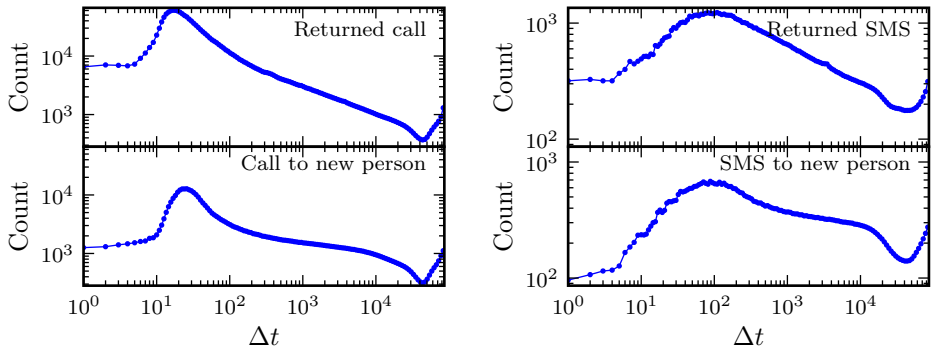
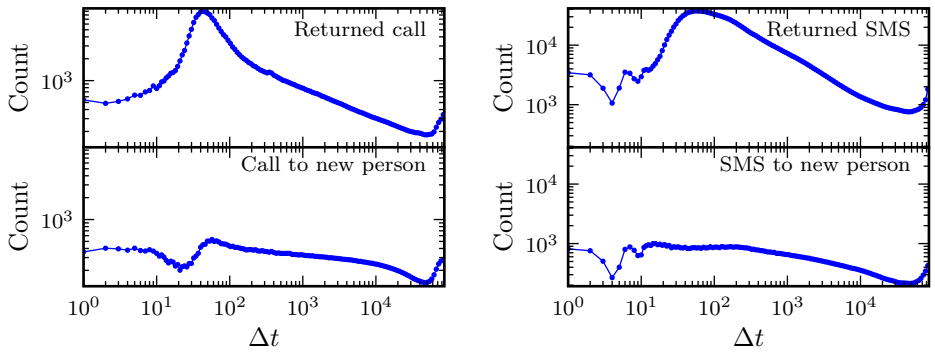


Figure 5.1: How action trigger plots are created. The numbers next to the edges (such as $0 \rightarrow 17$) tell the starting and ending times of each phone call. For instance in the uppermost graph A calls B at time 0 seconds and the call ends at time 17 s, after which B calls D at time 27 s. This results in one point in the action trigger plot at $\Delta t = 10s$, corresponding to the time it took B to call D after finishing the call with A. In the middle graph both A and E call C, after which C calls D, resulting in two additional points in the action trigger plot. Going through all outgoing calls, and all incoming calls before each outgoing call, we get the complete action trigger plot (bottom).



(a) From call to call

(b) From call to SMS



(c) From SMS to call

(d) From SMS to SMS

Figure 5.2: Action trigger plots with a time window of 24 hours. The horizontal axis is the time in seconds from the end of an incoming event (a call or an SMS message) to the beginning of an outgoing event. The duration of an SMS message is zero for all other than multipart messages, where the duration corresponds to the time difference between the first and the last parts. The vertical axis shows the total number of incoming events at any time. To reduce noise, the values have been averaged with logarithmic bin sizes, except for the first 10 seconds.

takes to mediate an SMS, and during this time there are only correlated, non-causal calls. It then takes the receiver approximately 20 seconds to read the SMS, and the maximum of returned calls is reached roughly one minute after the SMS was sent. Return calls after receiving an SMS message show a very different behaviour, but comparing the upper plots in Figures 5.2(a) and 5.2(c) we can see that it typically takes 24 seconds longer to respond to an SMS message than to a phone call — again, time needed to mediate and read an SMS message.

The time taken to write an SMS message may be seen in Figure 5.2(b). The number of returned SMS messages rises slowly from 5 second to a flat peak with a maximum at 100 seconds, reflecting the varying length of SMS messages and the time it takes to write one. In contrast with returned calls in Figure 5.2(a) it takes only 17 seconds to reach the maximum.

5.2 References

Of course, not all calls made after an incoming call can be said to be *caused* by that call. We need to take a closer look to differentiate causality from correlation, and to explain the shape of the action trigger in general.

The simplest possible way to calculate a reference would be to take the total number of (received) calls (83.8 million in the events data) and the number of seconds in one month (2.68 million) to conclude that if the calls were uniformly distributed on all users and over time, the action trigger plot would be flat with value 31.3. This reference clearly doesn't include any causalities, but as the true value is about 1000 times larger, we are obviously doing way too many approximations.

The difference between the naïve average value and the real value stems from several sources:

- Saying that all users make (or receive) roughly the same number of calls is just plain wrong. The distribution is fat-tailed; 12 % of the most active users make half of the calls, and the top 2 % make 15.8 % of all calls. (see Figure 3.4(b)).
- Events are not uniformly distributed in time, as is evident from Figures 3.5(a) and 3.6. Calls are strongly concentrated on few hours of the day, with a different pattern on weekdays than weekends.
- The temporal calling pattern of each individual is different from the average.

- There are external reasons for the correlation of calls. For example, a big party with friends is likely to induce calls among the friends at roughly the same time, but it would be wrong to say that the earlier calls cause the later ones (even though that will most likely happen too) — an external cause creates temporal correlations.
- Finally, a call can be triggered by a previous call, as we have already concluded from the action trigger plots.

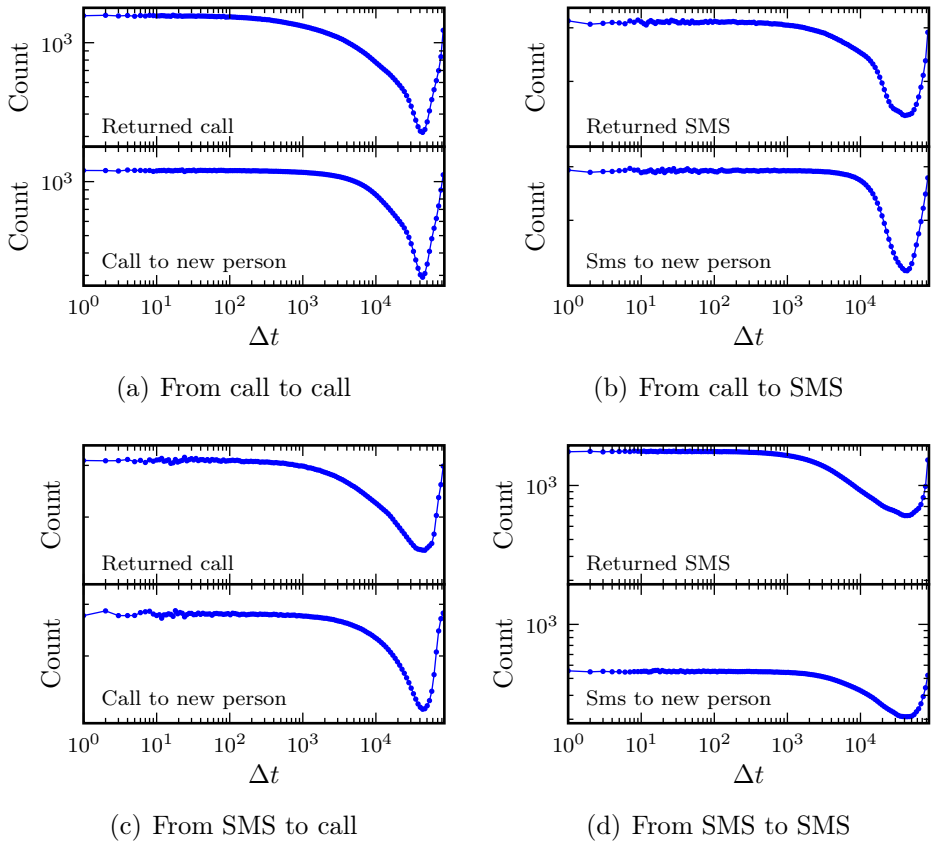
5.2.1 Reference as average over other days

The first non-trivial reference could be constructed by shuffling the calling times and calculating the action triggers for the obtained data. This would retain the strength of each user and the average temporal calling pattern and remove all traces of correlation and causality, but it would also destroy individual calling patterns.

We can do better than simply assume all users to have the same temporal pattern. We construct a reference by averaging the incoming calls over *all other days* than the one the outgoing call was made. For example, if a person made a call on January 7th at 15:32:14, to create the normal action trigger (with a 24 hour time window) we'd look at incoming calls ending between January 6th 15:32:13 and January 7th 15:32:13. Instead, we look at all other days[1] from 15:32:13 onwards until the same time on the following day, count all incoming calls and divide the total count by the number of days.

The resulting action trigger plots can be seen in Figure 5.3. The spikes we saw in the action triggers have disappeared, and the plots are flat except for the drop at 12 hours corresponding to the average circadian rhythm. Looking at calls, the flat part has about 1400 calls per second.

[1]There are 29 'other' days since we exclude January 1st due to its disparate statistics.



(a) From call to call

(b) From call to SMS

(c) From SMS to call

(d) From SMS to SMS

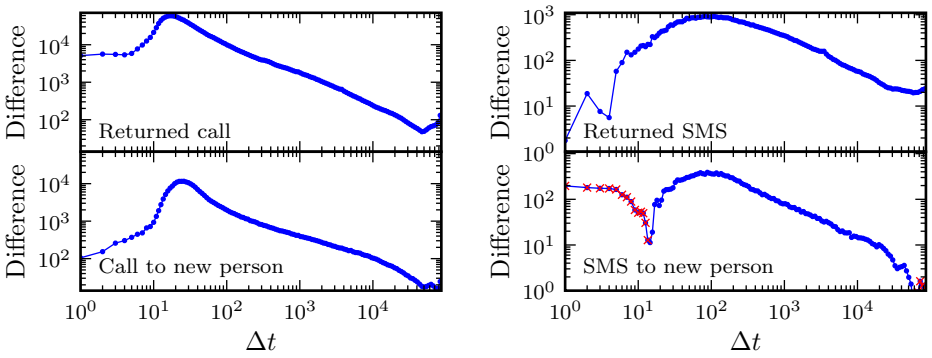
Figure 5.3: Action trigger plots when the calling times (and SMS message sending times) have been averaged over all other days.

5.2.2 Difference from reference

The next logical step is to see if the references are any good. Figure 5.4 shows the difference between the real action triggers and the references calculated above — all that is left consists of day-specific correlations and causalities.

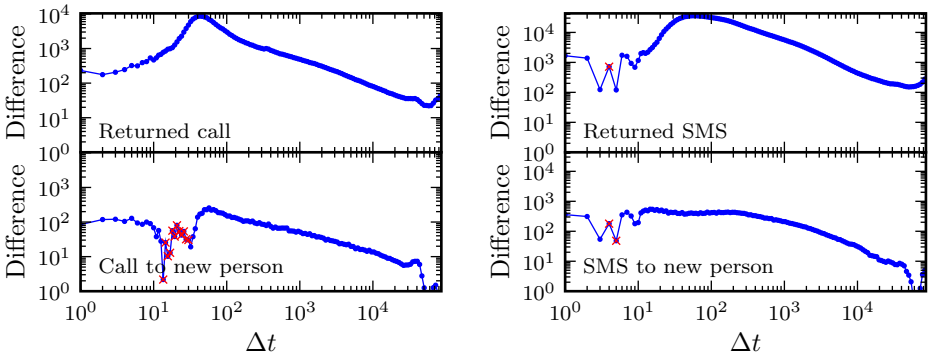
The thing to notice is that all graphs have roughly the same shape: after the initial rumble[2], each graph rises to a peak and then decreases roughly along to a straight line, hinting at a power law, until reaching the lowest point at roughly 12 hours.

[2]Probably caused by technical or practical reasons, such as the time taken to make a phone call or write an SMS message or the time needed to relay an SMS message.



(a) From call to call

(b) From call to SMS



(c) From SMS to call

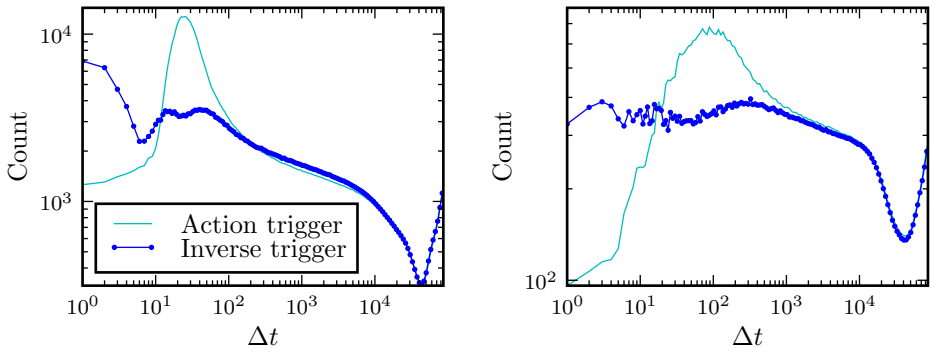
(d) From SMS to SMS

Figure 5.4: The difference between the true action trigger plots (Fig. 5.2) and the references (Fig. 5.3) in logarithmic coordinates. The red crosses denote negative values, but because it is not possible to show negative values with the logarithmic scale, absolute values are shown instead.

While we cannot tell whether the spike results from causality or correlation, it does seem that *the effect of the incoming call diminishes as a power law.*

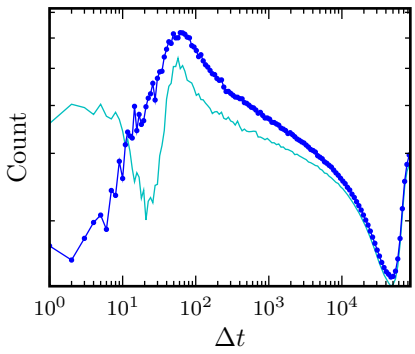
5.2.3 Inverse action trigger

So far we haven't been able to distinguish causality from correlation. But there is one blatantly obvious difference between the two: causality works in only one direction, while correlation should be identical no matter whether we look forwards or backwards in time. Using this idea we calculate *inverse action triggers*: instead of counting the number of incoming calls before an

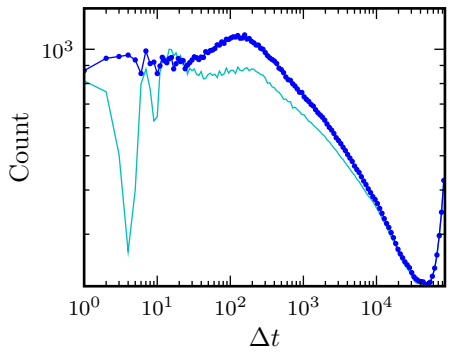


(a) From call to call

(b) From call to SMS



(c) From SMS to call



(d) From SMS to SMS

Figure 5.5: Inverse action trigger plots with a time window of 24 hours. The horizontal axis is the time in seconds from the end of an *outgoing* event (a call or an SMS message) to the beginning of an *incoming* event. The vertical axis shows the total number of incoming events at each time. To reduce noise, the values have been averaged with logarithmic bin sizes, except for the first 10 seconds where each time is represented separately.

outgoing call, we count the number of incoming calls *after* an outgoing call. The plots now show how long it takes to receive a call after making one call.

The inverse trigger plots for returned calls are identical to the corresponding action trigger plots, and the reason is explained in Figure 5.6(a).

The inverse triggers of new calls are shown in 5.5. We can see that even the inverse trigger plots have spikes. One possible explanation for this is causality mediated by a third party, illustrated in Figure 5.5. This would cause attenuated spikes (only calls cause a mediated causality) with a maximum

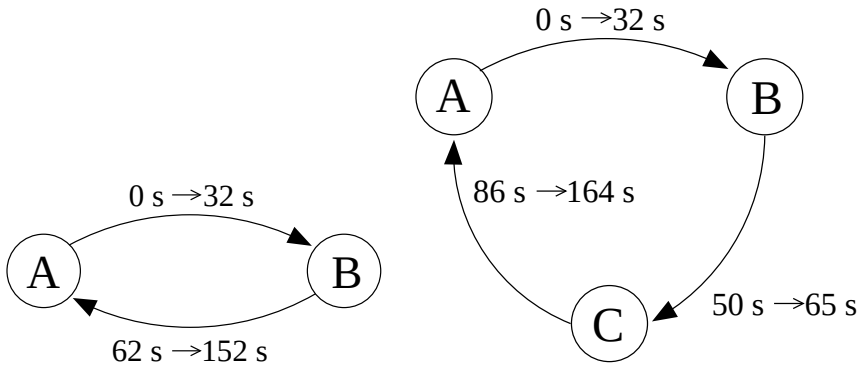


Figure 5.6: Nodes A, B and C correspond to people, edge labels (e.g. 0 s → 32 s) tell the starting time (0 s) and the ending time (32 s) of the calls. **(a)** Why inverse trigger and action trigger plots are identical for returned calls: To create the action trigger plot we search for incoming calls before an outgoing call, an thus there will be one point at $62 - 32 = 30$ s. With the inverse action trigger we look forwards in search for incoming calls, which again results in a point at time 30 s. **(b)** Why the inverse action trigger should also have a spike: The action trigger plot for this figure will have two points, one corresponding to the reaction of B (18 s) and another corresponding to the reaction of C (21 s). The inverse action trigger will have only one point corresponding to the full cycle starting and ending at A ($86 - 32 = 54$ s).

at roughly twice the time of the action trigger maximum, as is observed in Figures 5.5(a) and 5.5(b).

Surprisingly the maxima in the inverse trigger plots in Figures 5.5(c) and 5.5(d) are in fact higher than in the action trigger plots and occurs at roughly the same time. The actual reason for this curious observation would necessitate further study. Irrespective of the actual explanation, it does appear that the inverse action triggers can not sufficiently differentiate between correlations and causality.

5.3 Discussion

In this chapter we defined *action triggers* and applied them to study the causality of mobile phone communication. The action triggers were seen to

exhibit a spiking behaviour, and the location of the peak was identified as the *characteristic time* of the corresponding incoming and outgoing communication times.

As is often the case, distinguishing causality from correlation turned out to be difficult. By subtracting a reference signal we were able to extract the part of the action trigger plot that depicts only the daily causalities and correlations. The plots were seen to descend roughly linearly in the loglog-plot after the maximum, which would mean that the influence of an incoming call decreases as a power law.

The last effort to tell apart causality and correlation was based on the obvious fact that causality can only work in one direction in time. It turned out that the inverse action triggers were much more complex than expected. They also showed a similar spike as the action triggers, which can be explained by causality relayed through a third party.

In general, all reasoning that includes the timing is hindered by the noise caused by the multiple features of mobile handsets, properties of the communication technology and differences in personal usage behaviour. For example, the small characteristic time of returned calls could be explained by the fact that many handsets allow fast calls to the most recently used phone number. The analysis of SMS messages suffers from the delays in relaying the messages; in fact, the official SMS specification does not even guarantee the delivery of the messages.

The analysis presented here could be refined by taking into account different calling patterns on different days (as shown in Fig. 3.6) when calculating the reference: instead of averaging the incoming calls over all other days, we could average over all other days with the same day of the week. The analysis would be more precise, but would require more data. With only 4 weeks of time stamped data treating each day of the week separately doesn't leave many days to calculate the reference from.

Chapter 6

Conclusions and future work

In this thesis two novel concepts were introduced to study the structure and dynamics of complex social networks: the *edge bias* to study the reciprocity of weighted directed edges and the *action trigger* for the study of causality of mobile phone communication.

It was seen that there is a large variation in the reciprocity of edges. Even when there are more that 50 calls between two people during the period of 18 weeks, it is very common that one of the two is responsible for over 80 % of the calls. This is not a simple consequence of people having widely differing rates of activity; instead, reciprocity appears to be a property of the relation itself. While the true reasons behind the large edge bias values remain unknown, it was seen that the person with a higher degree is on average more active, and even more so when the edge is more active.

By using action triggers it was shown that mobile phone communication indeed has a significant quantity of causality — there are more outgoing calls shortly after an incoming call than later on. The most likely time to place a call to a new person after receiving a call is about 25 seconds, and the probability of making a call decreases as power law thereafter. Exactly why this shape appears is yet unclear, but since people often rely on their memory to make calls, it could be that this shape represents the rate at which

people forget recent events.

The conclusion is obscured by the difficulty of distinguishing causality from correlation. While it is impossible to say whether any two phone calls are causal or just correlated, we might still be able to say something about the averages. Unfortunately, exploiting the simple fact that causality only works in one direction in time didn't work quite as well as planned.

6.1 Next steps

This thesis has only taken the first steps in the study of both reciprocity and causality of communication. There is large number of issues that still need to be researched, both to confirm and to extend the results presented.

6.1.1 Reciprocity

We found out that the degrees of the caller and the recipient affect the reciprocity of the communication. There are no doubt other explanations, and incorporating the demographic information into the analysis could give more hints about the cause.

Because the study in this thesis was conducted by using only one (albeit large) data set, the obvious thing the check is whether the observed phenomena can be found in other data sets. This line of study is hindered by the lack of weighted, directed data sets, but for example instant messaging data sets could be used.

Using other data sets could also help bring some light on the most interesting question concerning reciprocity: does the large variation of reciprocities extend to social networks other that communication networks? In other words, do human relations have the tendency to be biased irrespective of how we define the relation? If so, what effect does this have on flow of information,

formation of opinions or on the society as a whole?

6.1.2 Causality

As with reciprocity, the existence of causality in communication should be confirmed with other data sets. In addition to other mobile phone communication data, the analysis could also be carried out with email communication data to find out whether the result holds for an entirely different kind of medium.

In addition to trying out other data sets, different methods for identifying causality should be developed and applied. It would be interesting to see, for example, how closely the characteristic times match with different methods. Possible alternative approaches include *Granger causality* [25] and information theoretic methods [26]. Also, as discussed earlier, the method of action trigger could be improved if more data was available, as this would allow the calculation of more accurate references.

There are many other open questions regarding causality. Are some people more susceptible to causal behaviour? Is causality stronger at some time of the day than another? Is causality stronger among edges with frequent communication, or is it more common that causal communication takes place on rarely used edges? These questions might be more difficult to answer than it seems, since we are only able to identify causality in a statistical sense — claiming that any single call is causal is a completely different matter.

Bibliography

[1] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International World Wide Web Conference*, 2008.

[2] W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33:452–473, 1977.

[3] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20), 11 2002.

[4] Jari Saramäki, Mikko Kivelä, J.-P. Onnela, Kimmo Kaski, and János Kertész. Generalizations of the clustering coefficient to weighted complex networks. *ArXiv Condensed Matter e-prints*, August 2006.

[5] S. E. Ahnert and T. M. A. Fink. Clustering signatures classify directed networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 78(3), 2008.

[6] Sara N. Soffer and Alexei Vázquez. Network clustering coefficient without degree-correlation biases. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 71(5), 2005.

[7] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, February 2009.

[8] A. Arenas, L. Danon, A. Diaz-Guilera, P. Gleiser, and R. Guimera. Community analysis in social networks. *The European Physical Journal B - Condensed Matter*, 38(2):373–380, March 2004.

[9] M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, January 2001.

[10] Fredrik Liljeros, Christofer R. Edling, Luis A. Amaral, Eugene H. Stanley, and Yvonne Aberg. The web of human sexual contacts. *Nature*, 411 (6840):907–908, June 2001.

[11] Renaud Lambiotte, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, September 2008.

[12] J.-P. Onnela, J. Saramäki, Hyvönen J., Szabó G., Lazer D., Kaski K., Kertész J., and Barabási A.-L. Structure and tie strengths in mobile communication networks. *PNAS*, 104(18):7332–7336, May 2007.

[13] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, A.M. de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9, 2007.

[14] Gergely Palla, Albert-Laszlo Barabasi, and Tamas Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, April 2007.

[15] Petter Holme. Network reachability of real-world contact sequences. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 71 (4), 2005.

[16] Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network, Jun 2008.

[17] Vassilis Kostakos. Temporal graphs. *Physica A: Statistical Mechanics and its Applications*.

[18] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969. URL `http://www.jstor.org/stable/2786545`.

[19] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[20] E. Limpert, W. A. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352, May 2001. ISSN 0006-3568.

[21] Christian A. Silva and Victor M. Yakovenko. Temporal evolution of the "thermal" and "superthermal" income classes in the usa during 1983-2001. Oct 2004. URL `http://arxiv.org/abs/cond-mat/0406385`.

[22] Jukka-Pekka Onnela. *Complex networks in the study of financial and social systems*. PhD thesis, Helsinki University of Technology, 2006.

[23] M. E. J. Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3): 035101+, 2002.

[24] Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the neural code*. The MIT Press, 1999.

[25] Mingzhou Ding, Yonghong Chen, and Steven L. Bressler. Granger causality: Basic theory and application to neuroscience, Aug 2006.

[26] Milan Paluš and Aneta Stefanovska. Direction of coupling from phases of interacting oscillators: An information-theoretic approach. *Physical Review E*, 67(5):055201+, May 2003.